Title

Analyzing the Dynamics of Wildfires: Causes, Patterns, and Predictive Modeling of Large and Small Fires in the United States

Author

Wang, Zixuan

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Analyzing the Dynamics of Wildfires:
Causes, Patterns, and Predictive Modeling of Large and Small Fires in the United States

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics and Data Science

by

Zixuan Wang

2024

ABSTRACT OF THE THESIS

Analyzing the Dynamics of Wildfires:

Causes, Patterns, and Predictive Modeling of Large and Small Fires in the United States

by

Zixuan Wang

Master of Applied Statistics and Data Science

University of California, Los Angeles, 2024

Professor Yingnian Wu, Chair

This study aims to analyze the dynamics of wildfires in the United States by predicting fire size for both small and large fires and developing a classification model for fire size. Using a dataset spanning from 1992 to 2015, machine learning models such as XGBoost, CatBoost, Random Forest, Generalized Linear Models (GLMs), and Mult-layer Perceptron (MLP) were applied to predict fire size, with XGBoost and CatBoost showing strong performance in predicting small and large fires, respectively. Additionally, classification models, including XGBoost, CatBoost, and Random Forest, were developed to distinguish between small and large fires, with challenges arising from class imbalance. Future work will focus on improving model performance by incorporating more detailed environmental data and exploring advanced machine learning techniques.

The thesis of Zixuan Wang is approved.

Xiaowu Dai

Robert L. Gould

Maria Cha

Yingnian Wu, Committee Chair

University of California, Los Angeles

2024

To my family and friends, whose unwavering support and encouragement have been a constant source of strength throughout this journey.

TABLE OF CONTENTS

1 Introduction **1**

2 Literature Review **3**

2.1 Factors Influencing Wildfire Occurrence 3

2.2 Characteristics of Large Fires 4

2.3 Seasonal and Environmental Conditions 4

2.4 Predictive Modeling of Wildfires 5

2.5 Gaps in the Literature 6

3 Exploratory Data Analysis **7**

3.1 Overview of Wildfire Data 7

3.2 Analysis of Annual Number of Fires 10

3.3 Analysis of Annual Number of Fires by Cause 12

3.4 Analysis of Annual Fire Sizes 16

3.5 Geographical Factors 21

3.6 Analysis of the Reporting Agency & Introducing New Variable - Land Type 25

3.7 Analysis of the Average Temperature When Fires Occur 31

3.7.1 Average Temperature by State When Large Fires Occur 32

3.7.2 Average Temperature by State When Small Fires Occur 33

3.7.3 Average Temperature by Year When Large Fires Occur 34

3.7.4 Average Temperature by Year When Small Fires Occur 35

3.8 Correlation Heatmaps 35

4 Methodology 38

4.1 Log Transformation 38

4.2 Regression Metric: Root Mean Squared Error (RMSE) 41

4.3 Classification Metrics 42

4.4 Extreme Gradient Boosting (XGBoost) 45

4.5 Categorical Boosting (CatBoost) 46

4.6 Random Forest 49

4.7 Generalized Linear Models (GLMs) 51

4.8 Multi-layer Perceptron (MLP) 52

5 Results and Discussion 56

5.1 Small Fire Size Prediction 56

5.1.1 RMSE Results for Small Fire Size Prediction 56

5.1.2 Analysis and Comparison of Results 57

5.2 Large Fire Size Prediction 60

5.2.1 RMSE Results for Large Fire Size Prediction 60

5.2.2 Analysis and Comparison of Results 60

5.3 Fire Size Classification Model 61

5.3.1 Classification Results for Large Fire Prediction 62

5.3.2 ROC Curves for Large Fire Prediction 65

5.3.3 Analysis and Comparison of Classification Models 67

6 Limitation and Future Work 69

7 Conclusion 71

List of Figures

3.1 Annual Total Fire Size for Large Fires 10

3.2 Annual Total Fire Size for Small Fires 11

3.3 Total Annual Large Wildfires by Cause 12

3.4 Fire Causes Distribution for Large Fires 13

3.5 Total Annual Small Wildfires by Cause 14

3.6 Small Wildfires Cause Distribution 14

3.7 Fire Size by Year for Each State 16

3.8 Large Fire Size by Year for Each State 17

3.9 Alaska Fire Cause Distribution between 2004 and 2005 17

3.10 Alaska Large Fire Cause Distribution between 2004 and 2005 17

3.11 Small Fire Size by Year for Each State 18

3.12 Total Acres Burned Annually by Large Fires 19

3.13 Total Acres Burned Annually by Small Fires 19

3.14 Total Acres Burned Annually by Causes 20

3.15 Fire Location Heatmap in 1992 21

3.16 Fire Location Heatmap in 2002 21

3.17 Fire Location Heatmap in 2014 22

3.18 Fire Location Heatmap for Small Fires in 2002 23

3.19 Fire Location Heatmap for Large Fires in 2000 23

3.20 Fire Location Heatmap for Large Fires in 2007 24

3.21 Fire Location Heatmap for Large Fires in 2011 24

3.22 Fire Count by Agency 25

3.23 Fire Location by Reporting Agency Map 26

3.24 Fire Count by Land Type 29

3.25 Large Fire Count by Land Type 30

3.26 Large Fire Count by Agency 30

3.27 Small Fire Count by Agency 31

3.28 Average Temperature by State When Large Fire Occurs 32

3.29 Average Temperature by State When Small Fire Occurs 33

3.30 Average Temperature by Year When Large Fire Occurs 34

3.31 Average Temperature by Year When Small Fire Occurs 35

3.32 Correlation Heatmap for Large Fires 36

3.33 Correlation Heatmap for Small Fires 36

4.1 Left: Original fire size distribution (Small); Right: Transformed fire size distribution after Log Transformation (Small). 40

4.2 Left: Original fire size distribution (Large); Right: Transformed fire size distribution after Log Transformation (Large). 40

4.3 Representation of the MLP architecture with input, hidden, and output layers (Source: Chan et al., 2023 [4]). 54

5.1 GLM with Log Transformation Q-Q Plot 58

5.2 GLM with Log Transformation Residual Histogram 58

5.3 GLM with Log Transformation Residual Plot 59

5.4 GLM with Log Transformation Q-Q Plot 61

5.5 GLM with Log Transformation Residual Histogram 62

5.6 GLM with Log Transformation Residual Plot 62

5.7 ROC Curve for CatBoost (with class weights) for Large Fire Prediction 66

5.8 ROC Curve for XGBoost for Large Fire Prediction 66

5.9 ROC Curve for Random Forest for Large Fire Prediction 67

List of Tables

3.1 Summary of Wildfire Data (General) 8

3.2 Summary of Small Fires (Fires $\leq$ 1000 acres) 9

3.3 Summary of Large Fires (Fires $>$ 1000 acres) 9

5.1 RMSE values for predicting small fire size before and after applying log transformation. 57

5.2 RMSE Results for Large Fire Size Prediction before and after Log Transformation 60

5.3 Classification Report for CatBoost (without class weights) 63

5.4 Additional Metrics for CatBoost (without class weights) 63

5.5 Classification Report for CatBoost (with class weights) 63

5.6 Additional Metrics for CatBoost (with class weights) 64

5.7 Classification Report for XGBoost 64

5.8 Additional Metrics for XGBoost 64

5.9 Classification Report for Random Forest 65

5.10 Additional Metrics for Random Forest 65

ACKNOWLEDGMENTS

I would like to extend my deepest gratitude to my committee members for their invaluable contributions to the successful completion of this thesis.

First and foremost, I sincerely thank Dr. Yingnian Wu, my committee chair, for his unwavering support, insightful guidance, and encouragement throughout this journey. His expertise and mentorship have been pivotal in shaping the direction of this work.

I am also profoundly grateful to Dr. Maria Cha and Dr. Xiaowu Dai for their constructive comments and approval of this thesis. Their thoughtful feedback has significantly enriched the quality of this work.

A special thanks to Dr. Robert L. Gould for his detailed and thoughtful feedback. His comments and encouragement have provided both direction and motivation, for which I am truly appreciative.

Finally, I would like to express my heartfelt appreciation to my family for their endless love, encouragement, and support. Their belief in me has been my source of strength. I also extend my gratitude to my loyal and loving dog, who has been a constant companion and source of comfort during this journey.

This thesis would not have been possible without the contributions and support of all these individuals, and I am truly grateful for their presence in my life.

CHAPTER 1

Introduction

Wildfires are among the most dangerous natural hazards, posing significant risks to both human populations and ecosystems. Recently, the frequency and intensity of wildfires have increased dramatically, driven by factors such as climate change, land management practices, and human activities [1]. While the immediate impacts on communities and ecosystems are evident, the long-term consequences include air quality degradation, habitat destruction, and substantial economic losses. Effectively addressing these challenges requires a comprehensive understanding of wildfire causes, emerging trends, and the differences between large and small fires. Such knowledge forms the foundation for effective wildfire management and mitigation.

This study aims to explore the multifaceted nature of wildfires by examining the causes and characteristics of both large and small fires across the diverse landscape of the United States, incorporating data from all 50 states, the District of Columbia, and Puerto Rico.

The research will analyze large-scale spatiotemporal patterns of wildfires using a dataset spanning from 1992 to 2015, synthesized from federal, state, and local reporting systems. In addition, temperature data—including average monthly temperatures from U.S. states and territories—will be integrated to enhance understanding of the climatic influences on wildfire dynamics.

The research focus on Exploratory Data Analysis (EDA) to investigate the root causes and distinguishing characteristics of both small and large fires. Predictive modeling will primarily focus on small fires (less than 1,000 acres in size), while also exploring the causes of large fires and their relationships to land use changes, human activities, and climate-related variables. The ultimate goal is to provide actionable recommendations for improving fire management strategies.

To support this effort, machine learning techniques such as XGBoost, CatBoost, and Random Forest will be employed to develop predictive models assessing the likelihood of a fire being classified as large or small. These methods aim to identify the key factors driving large fires, offering insights that could inform more effective management approaches.

This study seeks to apply advanced methodologies, including EDA and cutting-edge machine learning models such as XGBoost, CatBoost, Random Forest, Generalized Linear Models (GLMs), and Multi-layer Perceptron (MLP). These approaches are designed to generate valuable insights that can improve wildfire management practices and strengthen community resilience.

The expected findings will contribute significantly to the ongoing discourse on wildfire management and resilience, providing practical implications for policies and strategies aimed at reducing wildfire risks across the United States.

CHAPTER 2

Literature Review

2.1 Factors Influencing Wildfire Occurrence

The dynamics of wildfire is due to the interaction of complex climatic, ecological, and anthropogenic factors. Indeed, Abatzoglou and Williams (2016) documented that climate change in the Western United States induces wildfires that are more frequent and larger. Their research identifies temperature and moisture as some of the leading variables that control the behavior of fire, arguing that anthropogenic climate warming has elevated fuel aridity and subsequently enhanced the likelihood of wildfire occurrences. This realization confirms that this research needs to capture the average monthly temperatures since such information helps identify with precision how climate variability alters fire behaviors towards climate-sensitive management of nature [1].

Amatulli et al. (2007) point out to natural and human-induced causes as the ignition sources of wildfires. The study indicates that such small-scale fires are more frequently caused by human activities involving objects such as cigarette butts and/or unattended campfires. Thus, knowledge of the particular contexts in which these events take place is crucial to EDA in the small fire data analyzed here, in that it provides a framework for classification and investigation of various causes of ignition [2].

This solution will give the general direction towards a good and meaningful report. Similarly, Radeloff et al. (2018) presented the recent trend of wildland-urban interface (WUI) across the United States. This is because the expansion of urban areas into more and more previously undeveloped areas increases the ignition sources and the difficulties with fire suppression. Their work points to the need, when considering the risk related to wildfires, to integrate urbanization as an important part of their analysis, seeing that human development continues to increase its role in fire dynamics in fire-prone regions [11].

2.2 Characteristics of Large Fires

Large wildfires have different characteristics from ordinary fires. San-Miguel et al. (2020) studied how different ecological factors, such as topography and fuel types, drive the behavior of fires in the boreal forests. In their work, it has been supported that such knowledge of these environmental controls is essential in the prediction of large fires, which will also be discussed in the exploratory analysis of this study [14].

Moreover, Radeloff et al. (2018) point out the significant role of urban expansion in amplifying the risk and impact of large fires. As housing continues to encroach on fire-prone areas, fire management becomes increasingly complicated. Their study highlights the complex interaction between urbanization and large fire occurrences, providing critical insights into the socio-economic factors contributing to the growing wildfire risk [11].

2.3 Seasonal and Environmental Conditions

As an insight, the understanding of wildfire occurrence with seasonal variations might provide crucial knowledge about the dynamic features of fire. Alex et al., 2023, conducted seasonal trend analysis in fire occurrence across Southern Africa, using the BFAST model - Breaks for Additive Season and Trend. From this study, it can be noted that seasonal cycles are a significant contributor to the frequency of fires and variations in intensity, which is a very important knowledge contribution regarding the large and small size variations in wildfires due to seasonal factors [8].

2.4 Predictive Modeling of Wildfires

Machine learning techniques have become an integral part of wildfire prediction. Bjånes et al. (2021) proposed a deep learning ensemble model to assess wildfire susceptibility, demonstrating how combining multiple deep learning algorithms can improve predictive performance. This study supports the use of advanced models, such as the ones explored in this research, to enhance wildfire prediction and risk mapping [3].

Pham et al. (2020) conducted a comprehensive performance evaluation of machine learning algorithms, including Random Forest and Support Vector Machines, for forest fire modeling and prediction. Their results indicated that machine learning methods outperform traditional statistical techniques in predictive accuracy, thus supporting the decision to incorporate various machine learning algorithms, including CatBoost, into the current study's methodology [10].

Rodrigues et al. (2014) examined machine learning algorithms developed solely for human-caused wildfire modeling. The comparative analysis done demonstrated that methods such as Random Forests and Boosted Regression Trees resulted in much better predictive performance than their predecessors, including logistic regression. This research further justifies the use of cutting-edge algorithms in the present study, particularly when exploring both large and small wildfire predictions [13].

In addition, Jiménez-Ruano et al. (2022) focused on spatial predictions of wildfire likelihood using logistic generalized linear models (GLMs) and generalized additive models (GAMs). Their research demonstrates the value of spatial modeling in understanding where wildfires are most likely to occur based on various environmental factors. These insights will guide the spatial modeling approach used in this study [6].

In this study, machine learning models, including CatBoost, XGBoost, Random Forest, GLMs, and MLP, will be utilized not only to predict wildfire size but also to predict the likelihood that a fire will be classified as large (greater than 1000 acres). This task aims to develop a classifier that assesses the potential size of a fire based on the same set of variables used in traditional wildfire predictions, such as fire causes, geographic coordinates, and seasonal trends.

2.5 Gaps in the Literature

While a substantial volume of research has been conducted on the factors influencing wildfires, significant gaps remain in understanding the comparative dynamics of large and small fires. Most prior studies have focused on overarching trends and causes, but the unique characteristics of small fires are still underexplored. Furthermore, while climate factors such as temperature are obviously important, they have not been fully integrated into modeling of wildfires.

Among these, the size classification of wildfires is of acute importance regarding the understanding of different types of dynamics that exist regarding fire behavior and management. Whereas large-size wildfires are normally those greater than 50,000 acres and medium-sized wildfires those between 1,000 and 50,000 acres, in general, small wildfires involve less than 1,000 acres. The threshold for a small fire has been widely taken on in understanding fire behaviors and management strategies, accordingly, in areas where wildfires contribute to increasing the risks through frequency and increasing scales. According to the ArcGIS Wildfire Size Interpretation, large Wildfires are more common than historically contributes a lot more to annual acres burned. The frequency and intensity of large fires have been increasing, especially in recent decades [5].

This research will fill these gaps through a comprehensive analysis of major and minor wildfires in the country. Using a robust dataset of wildfires across the country from 1992 to 2015, with temperature data integrated into the analysis, this study develops new insights into how climate, land use changes, and human activities shape the behavior of wildfires. The approach of the study covers not only causes but also behaviors of small fires, underlining how focusing on predictive modeling for smaller fires could set up new fruitful steps with regards to small and large fire management strategies.

CHAPTER 3

Exploratory Data Analysis

It is essential to conduct an in-depth exploratory analysis of the wildfire dataset to uncover key insights regarding fire sizes, trends, and the distribution of fire causes. Such an analysis provides a clearer understanding of the factors contributing to the occurrence and intensity of wildfires. This chapter focuses on examining the total number of wildfires, the total acres burned by wildfires, comparing small and large fires, and identifying the most frequent causes of these fires, followed by a case study. Additionally, temporal and geographical trends in wildfire occurrence are explored, as well as their relationship with fire size and frequency. By analyzing these factors, this chapter aims to highlight significant patterns and inform strategies for wildfire prevention and management.

3.1 Overview of Wildfire Data

The dataset comprises a total of 1,880,465 wildfires, which together have burned a total of 140,132,550 acres. Of these fires, the vast majority are small fires, with 1,869,378 fires having burned fewer than 1000 acres. These small fires account for 20,681,650 acres burned. The average fire size in the entire dataset is 74.52 acres; however, this distribution is skewed by a few extremely large fires, which is why the median fire size is only 1 acre. The smallest recorded fire is an incredibly small 1e-05 acres, while the largest fire measures an enormous 606,945 acres. Despite the large number of small fires, large fires, defined as those that exceed 1000 acres, account for a substantial portion of the total area burned. There are 11,087 large fires in the dataset, which together have consumed 119,450,900 acres. These large fires have an average size of 10,773.96

acres and a median size of 2,930 acres. The largest of these fires, which burned 606,945 acres, represents a significant portion of the total area burned by large fires.

Lightning is the most common cause of both small and large fires, followed by miscellaneous causes, debris burning, and arson. These facts highlight the significant role of natural factors like lightning in wildfire occurrences. Geographically, California (CA) stands out as the state with the highest number of wildfires overall, including both small and large fires. This is consistent with the state's vast size and diverse ecosystems, which are prone to frequent fires. Arizona (AZ) is notable for having the most large fires, further emphasizing the variation in wildfire patterns across different regions. The geographic distribution of fires demonstrates how environmental conditions, land use, and other factors contribute to the likelihood of fires in different parts of the country.

The summary statistics for the total dataset, small fires, and large fires are provided in the tables below.

Table 3.1: Summary of Wildfire Data (General)

Statistic	**Value**
Total number of fires	1,880,465
Total acres burned	140,132,550 acres
Average fire size	74.52 acres
Median fire size	1 acre
Smallest fire size	1e-05 acres
Largest fire size	606,945 acres
Most common cause of fires	Lightning
State with the most fires	California (CA)

Table 3.2: Summary of Small Fires (Fires $\leq$ 1000 acres)

Statistic	**Value**
Total number of small fires	1,869,378
Total acres burned by small fires	20,681,650 acres
Average small fire size	11.06 acres
Median small fire size	1 acre
Smallest small fire size	1e-05 acres
Largest small fire size	1,000 acres
Most common cause of small fires	Lightning
State with the most small fires	California (CA)

Table 3.3: Summary of Large Fires (Fires $>$ 1000 acres)

Statistic	**Value**
Total number of large fires	11,087
Total acres burned by large fires	119,450,900 acres
Average large fire size	10,773.96 acres
Median large fire size	2,930 acres
Smallest large fire size	1,000.1 acres
Largest large fire size	606,945 acres
Most common cause of large fires	Lightning
State with the most large fires	Arizona (AZ)

3.2 Analysis of Annual Number of Fires

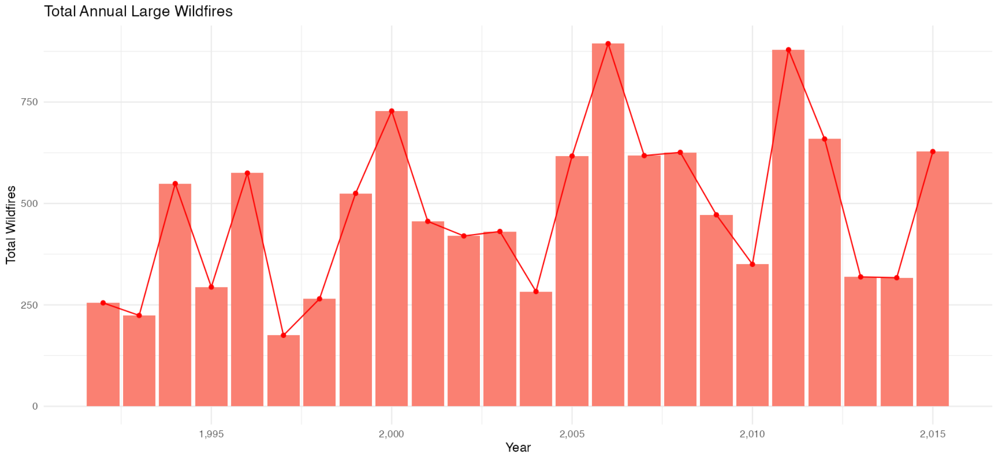

Figure 3.1: Annual Total Fire Size for Large Fires

Large fires, those over 1000 acres, vary significantly from year to year (Figure 3.1). From 1992 until the early 2000s, the number of large fires varied moderately, with several notable peaks in 1994, 1996, and 2000. Starting around 2005, the number of large fires increased sharply followed by a strong peak in 2006. After 2006 the number of large fires dropped, but then rose again sharply in 2011.

This could be due to several factors, which all summed up in 2006, including extreme weather conditions such as continued drought and higher temperatures. The 2006 peak might indicate an exceptional year of acute fire conditions promoted by environmental factors or human activities. Since this peak, the trend has been very variable; a slight decline post-2008 may reflect improved fire management or favorable weather, but the continued rise in large fires since 2010 reflects continued challenge to these fires, especially in areas prone to extreme fire potential.

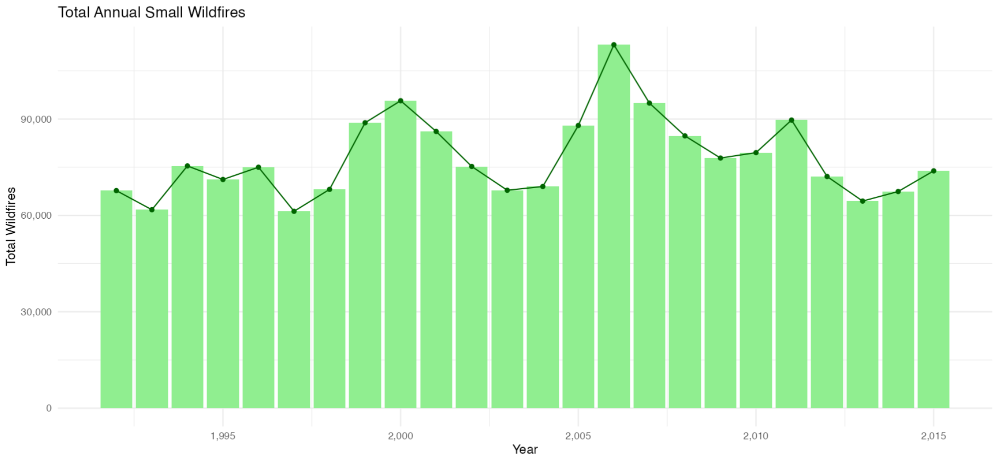

Figure 3.2: Annual Total Fire Size for Small Fires

In contrast, the graph for small fires (fires smaller than 1000 acres) shows a more consistent pattern over the years (Figure 3.2). Small fires had relatively stayed on the same level, although moderate oscillations could be observed for year-to-year variations. During the early 1990s to the mid-2000s, small fires were consistently frequent, with a slight downward trend emerging around the early 2000s. The peak of small fires occurred in 2006, after which there was a gradual decline in the following years. This does, apparently, reflect variability, yet large fires are outnumbered by small ones in terms of frequency. The trend in this graph illustrates that although large fires always will make up the greater portion of the landscape annually in acres burned, the number of small fires constitutes a significant component in wildfire occurrence. The persistence of small fires points to the omnipresence of the phenomenon, even if its impact remains limited by the area burnt compared with larger fires.

3.3 Analysis of Annual Number of Fires by Cause

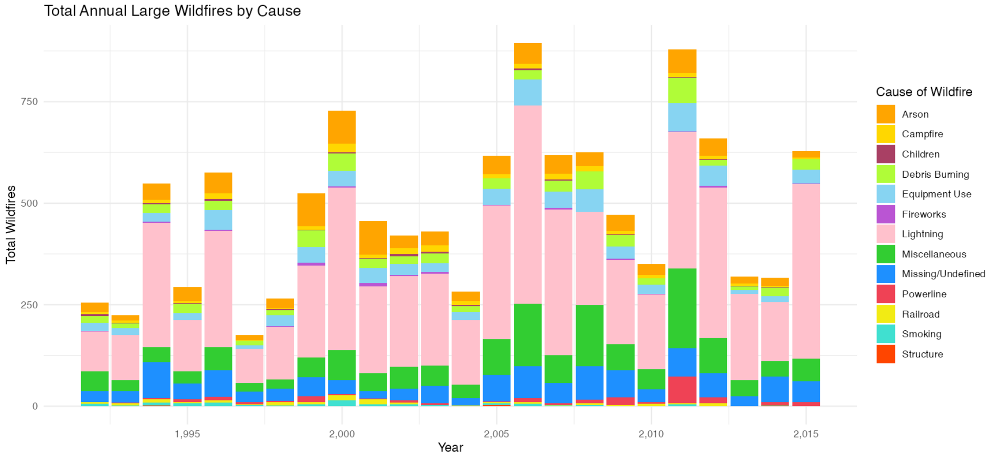

Figure 3.3: Total Annual Large Wildfires by Cause

The annual count of large wildfires shows distinct fluctuations based on the cause of the fire (Figure 3.3). **Lightning** is the consistent cause of large fires year after year, especially throughout the 2000s. Large wildfires due to lightning reach sharp peaks in 2000, 2006, and 2015, which most likely correlate with dry weather conditions and the increased number of lightning strikes during these years. Generally, lightning-induced wildfires are most influenced by environmental conditions, particularly vegetation moisture content. Dry lightning is one of the major natural ignition sources for wildfires and occurs with little or no precipitation. When the vegetation is dry, as in this case, the probability of such lightning strikes starting fires increases, since the vegetation is more prone to catching fire.

A study by Vecín-Arias et al. (2016) found that the occurrence of lightning-induced fires in Spain was highly influenced by the biophysical characteristics of the landscape, including vegetation type, weather, and lightning characteristics. The study highlighted that coniferous woodlands, which have highly flammable vegetation, are more prone to lightning-induced fires. This is consistent with the observed peaks in lightning-driven fires in the dataset, particularly in years with dry lightning conditions, where vegetation moisture is low and fires are more likely to ignite [16].

Additionally, a study by Rao et al. (2022) found that lower live fuel moisture content (LFMC) values, which correspond to drier vegetation, significantly increase the probability of lightning-induced wildfires [12]. Similarly, Kalashnikov et al. (2023) demonstrated that lightning-ignited wildfires can still occur under wetter conditions, provided that the vegetation is sufficiently dry, challenging previous assumptions that rainfall would prevent lightning from igniting fires. This finding suggests that the actual risk of fire ignition is influenced not only by increased lightning strikes but also by the moisture content of vegetation [7].

These studies emphasize the importance of considering both lightning activity and vegetation moisture content when assessing the risk of lightning-induced wildfires. Increased lightning strikes may elevate the potential for fire ignition, but the actual risk is greatly influenced by the dryness of the vegetation, which can vary due to factors such as drought, seasonal changes, and broader climate conditions.

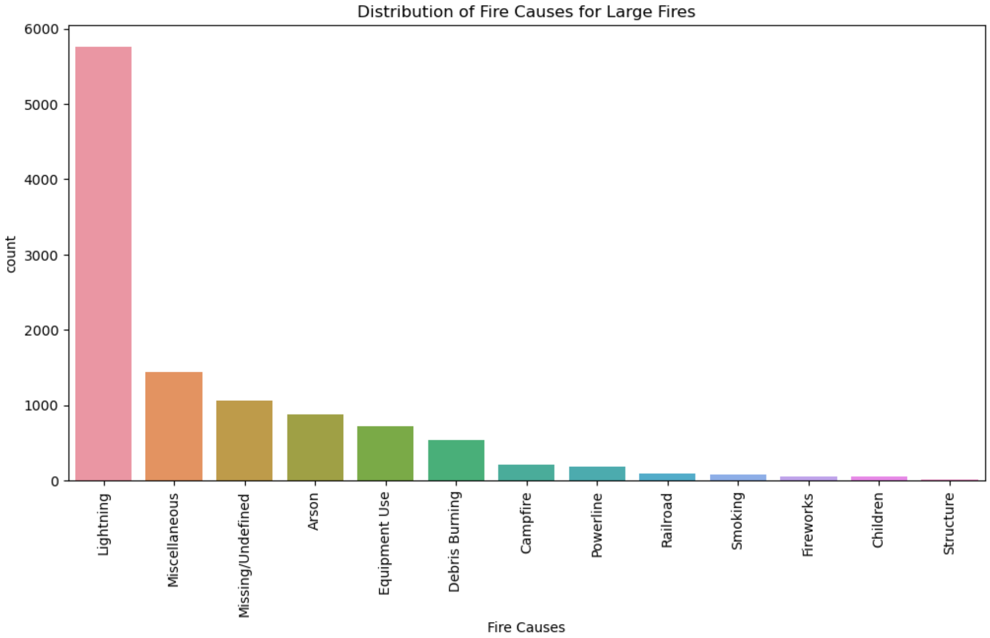

Figure 3.4: Fire Causes Distribution for Large Fires

Besides miscellaneous and undefined causes, **arson** is also a notable cause, contributing to several spikes in the data, particularly around the years 2000, 2006 and 2011 (Figure 3.4). Arson-related fires generally have a more irregular pattern compared to lightning-related fires, and the surge around 2006 could indicate a mix of human-caused fires alongside lightning-induced events, exacerbating the overall fire intensity. The presence of other causes, such as **debris burning** and **equipment use**, also contribute but to a lesser extent, with peaks occurring sporadically throughout the years.

These findings indicate that large wildfires are heavily influenced by both natural (lightning) and human-induced (arson, debris burning) factors. The sharp rise in large wildfires during specific years is often linked to a combination of severe weather conditions and human activity.

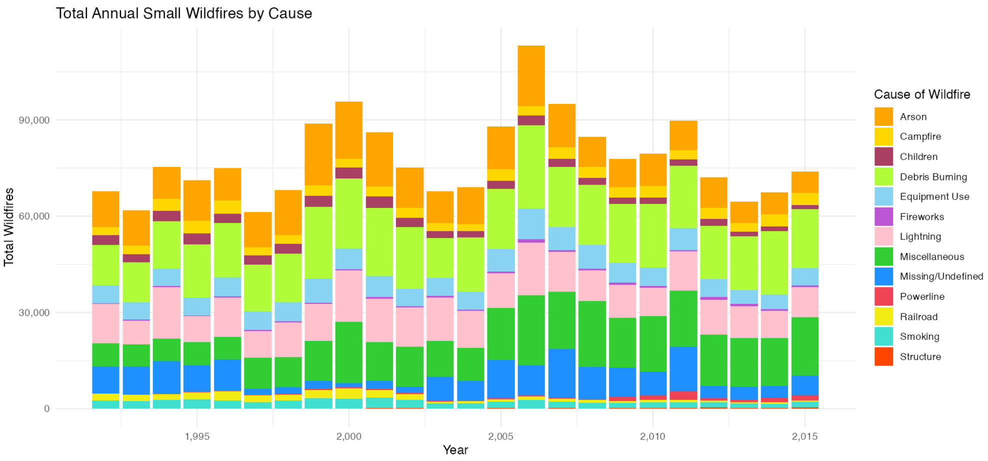

Figure 3.5: Total Annual Small Wildfires by Cause

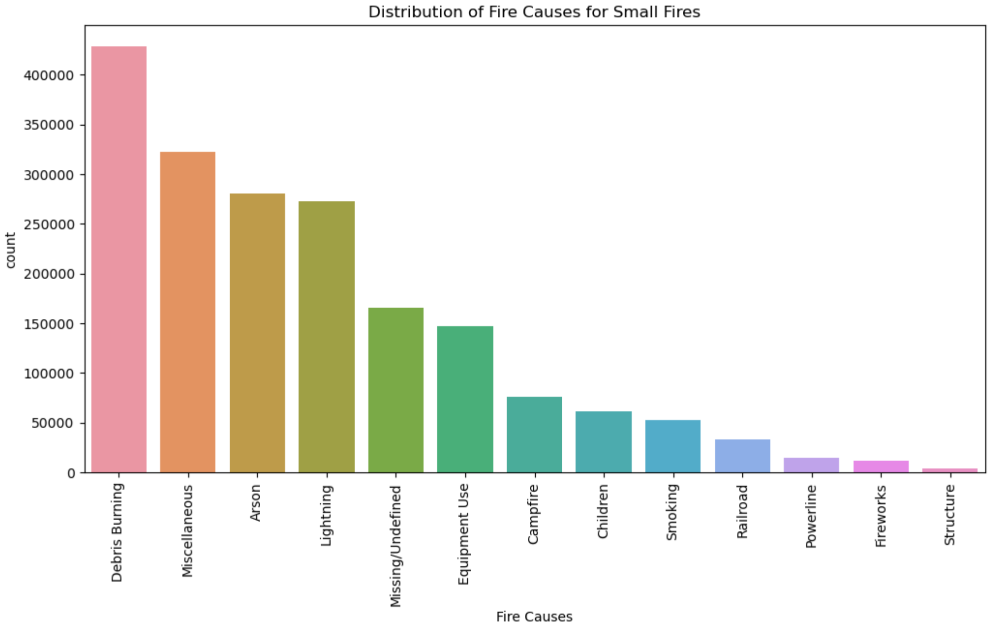

Figure 3.6: Small Wildfires Cause Distribution

Small fires, which make up the majority of wildfire incidents annually, show a more consistent pattern of occurrence, although the causes exhibit distinct trends (Figure 3.5). The top three causes of small wildfires are **arson**, **debris burning**, and **lightning** (Figure 3.6). Arson and debris burning contribute most significantly to the total number of small fires, which are often human-induced. These causes reflect the frequency of smaller, more controlled fires, typically resulting from illegal activities or agricultural practices.

From the mid-1990s to the mid-2000s, the frequency of small fires caused by **arson** and **debris burning** remained relatively high, with a notable peak in 2006. The increase in **debris burning** during the early 2000s is likely linked to agricultural activities, land-clearing, and burning practices that were prevalent during this time. The rise in debris burning could also be associated with regulatory changes, land management practices, or increased human activity in fire-prone areas.

Although **lightning** remains a significant cause of small fires, it is less frequent than arson and debris burning. Despite the fluctuations, small fires remain consistently higher in number compared to large fires. This trend demonstrates the importance of addressing fire prevention measures, particularly for human-induced causes like arson and debris burning, which contribute substantially to the overall number of wildfires.

The persistence of small fires highlights the widespread nature of wildfire occurrences, even though their impact in terms of area burned remains limited compared to larger fires. This trend demonstrates the need for a comprehensive approach to wildfire management, focusing not only on mitigating the risks associated with large fires but also addressing the more frequent smaller fires that occur due to human activities.

3.4 Analysis of Annual Fire Sizes

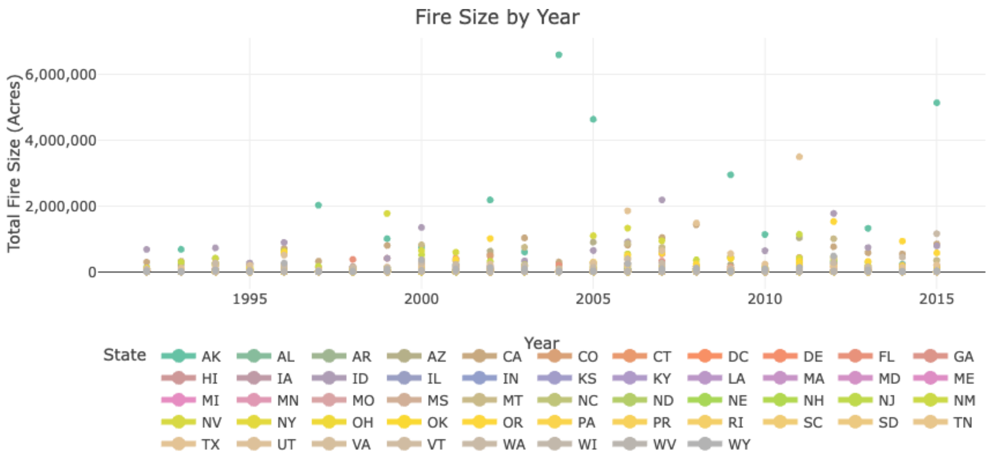

Figure 3.7: Fire Size by Year for Each State

The scatter plot depicting the total fire size by year for each state (Figure 3.7) illustrates the total acres burned by wildfires across different states over time. It is clear that some states, particularly Alaska (AK), Texas (TX), and Idaho (ID), exhibit substantial variations in fire size from year to year. These states tend to experience large fires with significantly higher total fire sizes compared to others.

From 1992 to the mid-2000s, fire sizes remained relatively stable across most states, with occasional spikes in Alaska and Nevada, where large fires are more common due to dry conditions and fire-prone landscapes. However, from 2005 onwards, several states, including Idaho and Texas, began experiencing noticeably larger fires, which contributed to the overall increase in total acres burned across the U.S. This trend is reflective of the growing intensity and frequency of wildfires in these regions, possibly exacerbated by climate change, prolonged droughts, and increased human activity.

The high variability in fire size across different states demonstrates the importance of considering regional factors, including climate, vegetation, and human activity, when analyzing wildfire risk and fire management strategies.

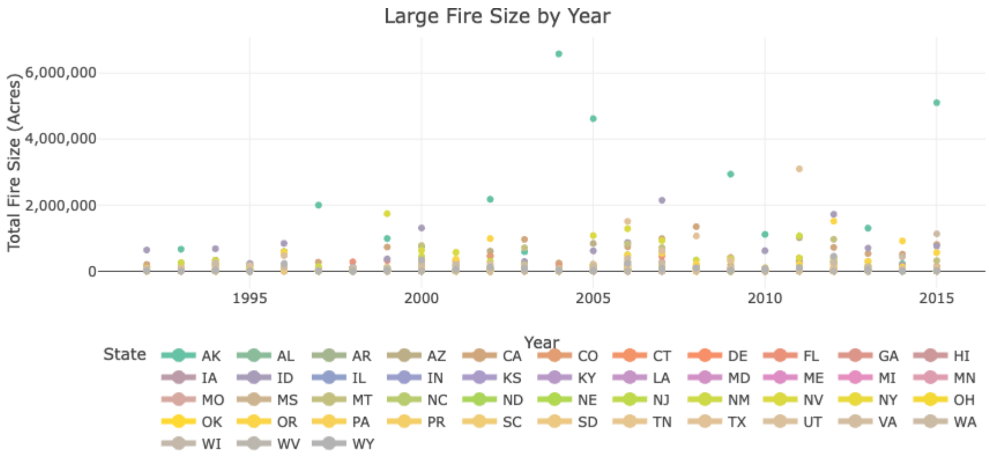

Figure 3.8: Large Fire Size by Year for Each State

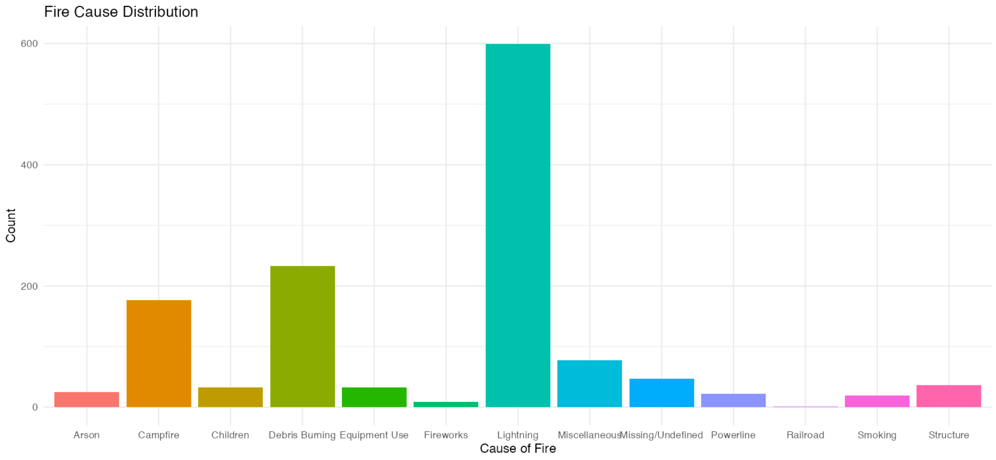

Figure 3.9: Alaska Fire Cause Distribution between 2004 and 2005

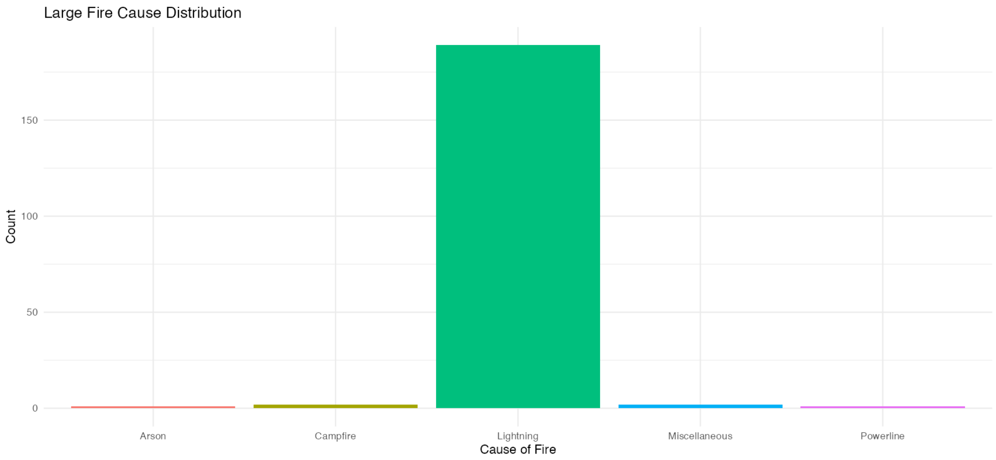

Figure 3.10: Alaska Large Fire Cause Distribution between 2004 and 2005

The scatter plot for large fires by year (Figure 3.8) specifically focuses on fires greater than 1000 acres. As expected, states like Alaska, Idaho, and Texas show the largest fire sizes, with substantial fluctuations observed each year. Notably, Alaska often experiences the highest fire sizes, particularly in years such as 2004, 2005, and 2015, which saw extreme fire seasons. Further analysis of the causes of Alaska's fires in 2004 and 2005 reveals that a significant portion of the fires were caused by lightning (Figure 3.9). Among large fires, nearly all of them are attributed to lightning (Figure 3.10). This illustrates the unique vulnerability of Alaska to lightning-induced fires, driven by its vast landscapes, frequent thunderstorms, and dry conditions during peak fire seasons. Given the geographical characteristics, where vast stretches of wilderness are often unreachable, Alaska's fire management strategies should prioritize early detection systems, lightning strike monitoring, and preparation for large-scale fire suppression. Additionally, improving land-use practices and firebreak strategies could help mitigate the frequency and intensity of such fires. The role of climate change in exacerbating the occurrence of extreme weather conditions, such as heatwaves and lightning storms, should also be factored into future fire prevention and response plans.

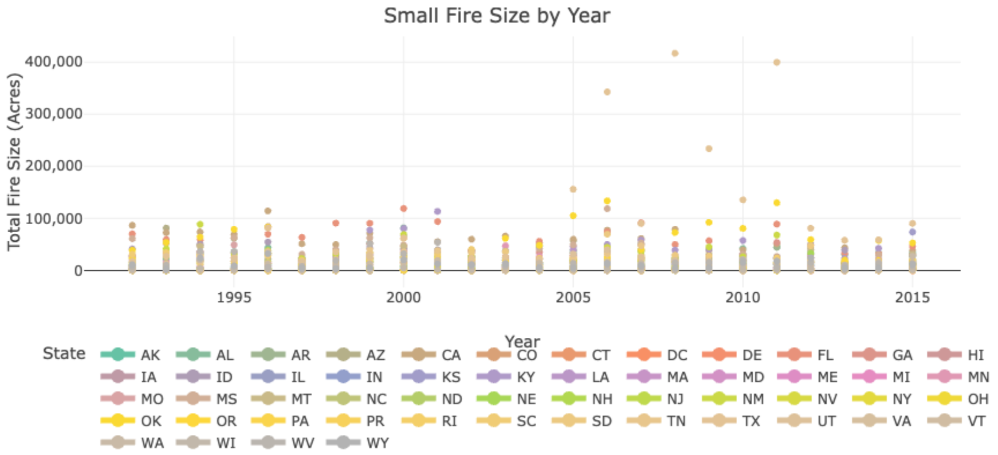

Figure 3.11: Small Fire Size by Year for Each State

In contrast, the scatter plot for small fires (Figure 3.11) shows less variation across states, but a consistent number of smaller fires occurring each year. Notably, Texas experienced a large fire size between 2005 and 2011, both in small and large fires. Small fires are widespread and occur annually in most states. While the total size of small fires is lower than that of large fires, their frequency means they still contribute significantly to the total number of fires in the dataset.

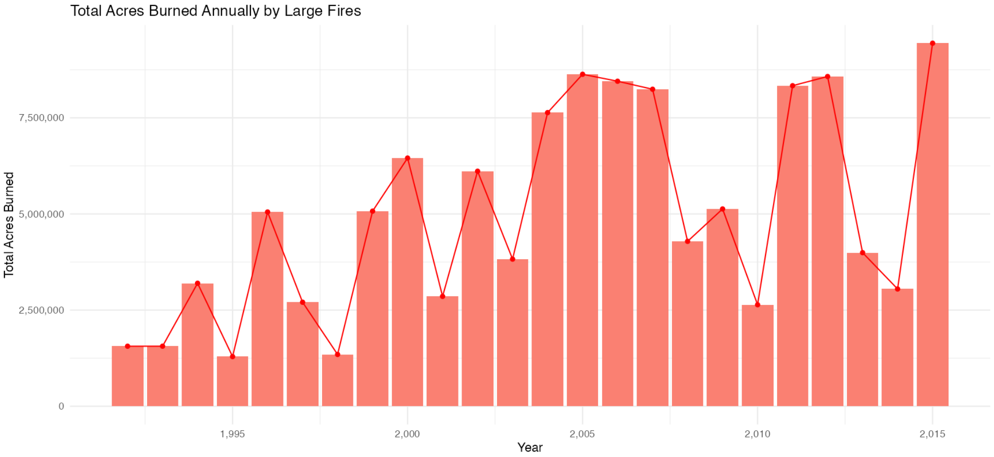

Figure 3.12: Total Acres Burned Annually by Large Fires

The total acres burned annually by large fires (Figure 3.12) reflects the high impact of large-scale fires on the total area burned each year. The graph shows distinct peaks in acres burned, particularly in 2005, 2006, 2007, 2011, 2012, and 2015, which coincide with years when extreme fire seasons were observed. The increase in total acres burned by large fires over time suggests that these fires are becoming more intense, contributing to a larger proportion of the total acreage affected by wildfires.

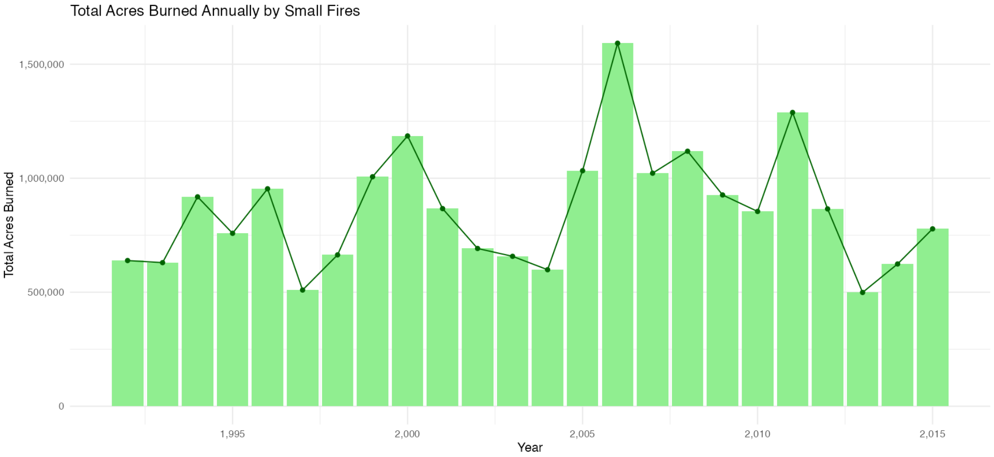

Figure 3.13: Total Acres Burned Annually by Small Fires

On the other hand, the total acres burned by small fires (Figure 3.13) show a more consistent trend, with only peaks in 2006 and 2011. While small fires burn a significant amount of land each year, they generally cover less acreage compared to large fires. Despite this, small fires are more numerous and frequent, contributing to a substantial portion of the total area burned by wildfires each year. These fires are often caused by human activities, such as arson, debris burning, and equipment use, and their persistence underscores the need for fire prevention measures, particularly in areas where small fires are prevalent.

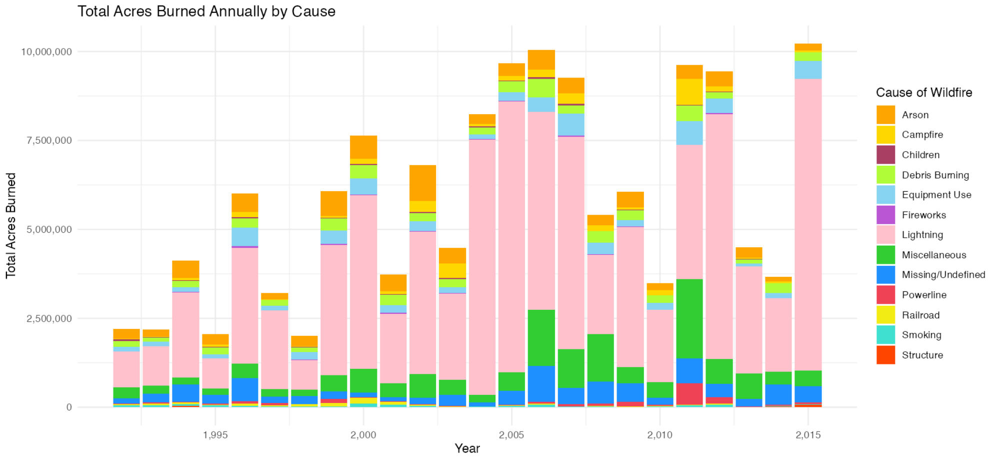

Figure 3.14: Total Acres Burned Annually by Causes

The stacked bar chart depicting total acres burned annually by cause (Figure 3.14) highlights the dominant role of human-induced factors in wildfire occurrences, especially lightning, miscellaneous causes, and arson. Arson consistently emerges as a leading cause, with notable spikes in several years, particularly around 2002, 1999, and 2000. Besides, the importance of lightning as a wildfire trigger becomes evident, particularly in the latter years, where dry weather conditions and lightning storms increased fire intensity. The steady contribution of miscellaneous and equipment use causes suggests that a variety of human activities, ranging from recreational activities to maintenance work, continue to pose significant wildfire risks.

3.5 Geographical Factors

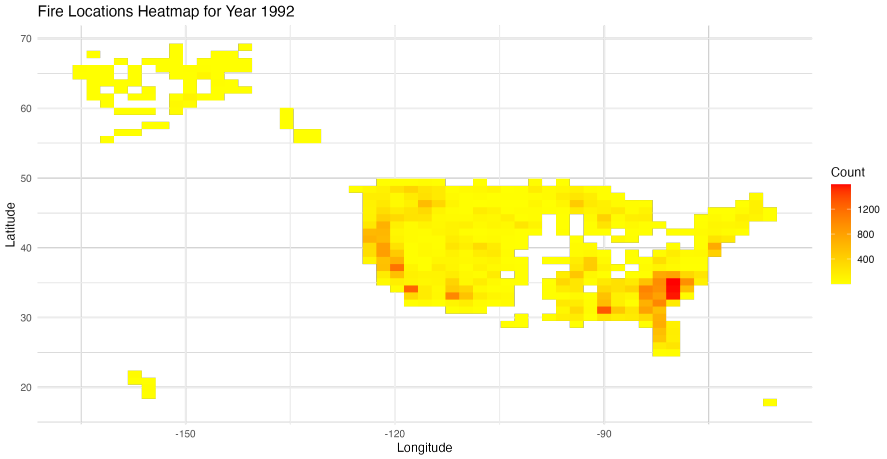

Figure 3.15: Fire Location Heatmap in 1992

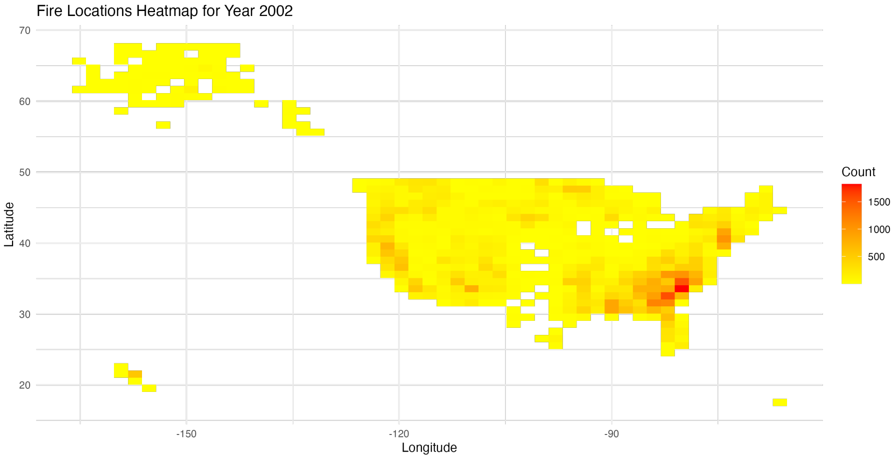

Figure 3.16: Fire Location Heatmap in 2002

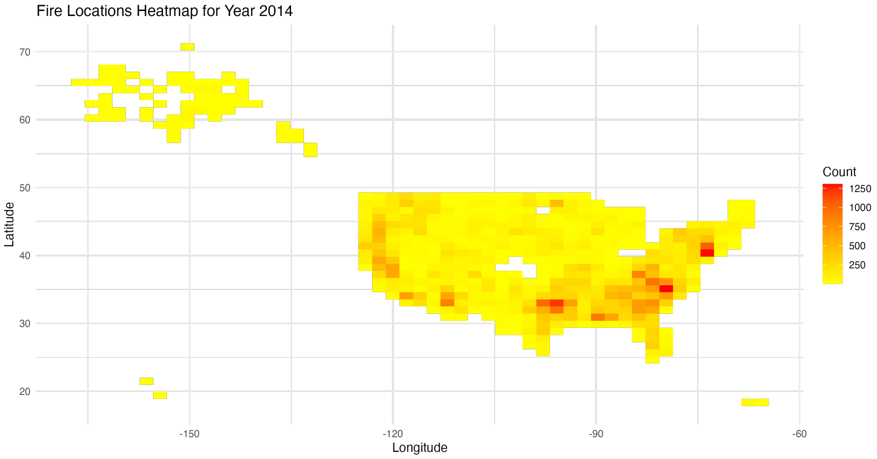

Figure 3.17: Fire Location Heatmap in 2014

Looking at the geographical location of the fires helps present a better view of an area that has a high frequency of fires. A series of fire location heatmaps ranging from 1992 to 2014 shows, for all these years, one consistent trend in the geographical locations, with the most intense activities happening in the southeastern United States and along the West Coast. While the intensity and frequency of fires fluctuate each year, this general pattern is noticeable throughout the time period. For instance, plots for 1992, 2002, and 2014 (Figures 3.15, 3.16, and 3.17) shows this distinct pattern.

In the example plots, the Southeast (including states like Florida and Georgia) consistently experiences a higher number of fires, likely due to favorable conditions for fire spread, such as the combination of seasonal humidity and temperature, alongside human activities like agriculture and land clearing. On the West Coast, California remains a persistent hotspot due to its dry vegetation, mountainous terrain, and exposure to strong winds during fire seasons, which often lead to more frequent wildfires.

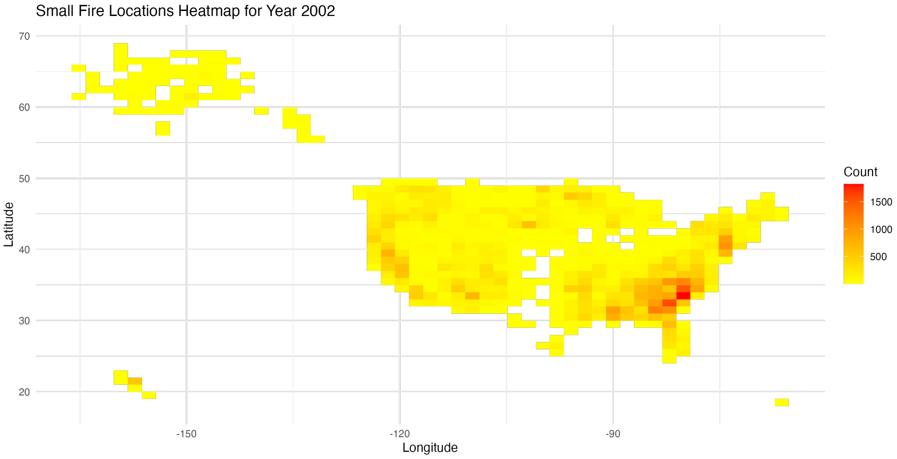

Figure 3.18: Fire Location Heatmap for Small Fires in 2002

As small fires make up the majority of wildfire incidents, the heatmap for small fires follows a similar geographical distribution, with areas in the Southeast and West Coast seeing the highest number of smaller fires. These regions, characterized by their dry vegetation and human activity, are highly susceptible to small fires, which can either result from natural causes like lightning or human-induced causes such as debris burning and arson. An example of small fires' location heatmap in 2002 is shown (Figure 3.18).

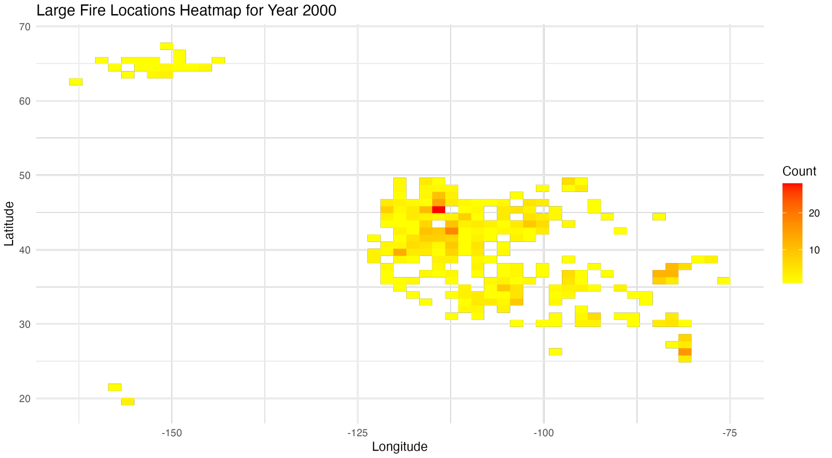

Figure 3.19: Fire Location Heatmap for Large Fires in 2000

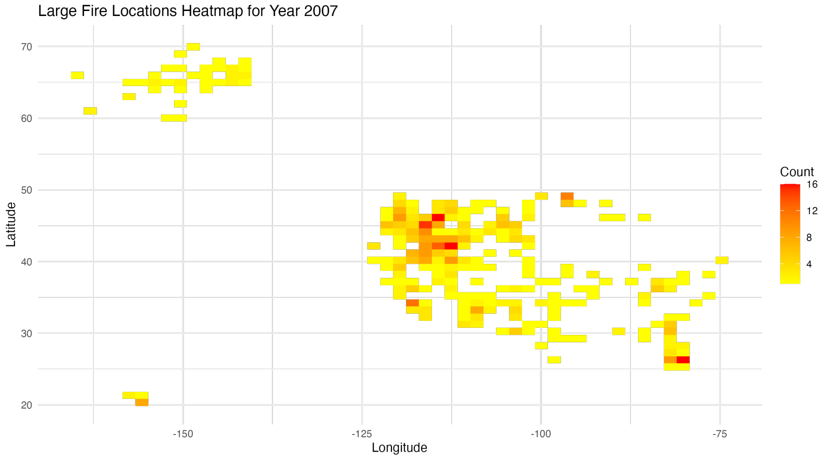

Figure 3.20: Fire Location Heatmap for Large Fires in 2007

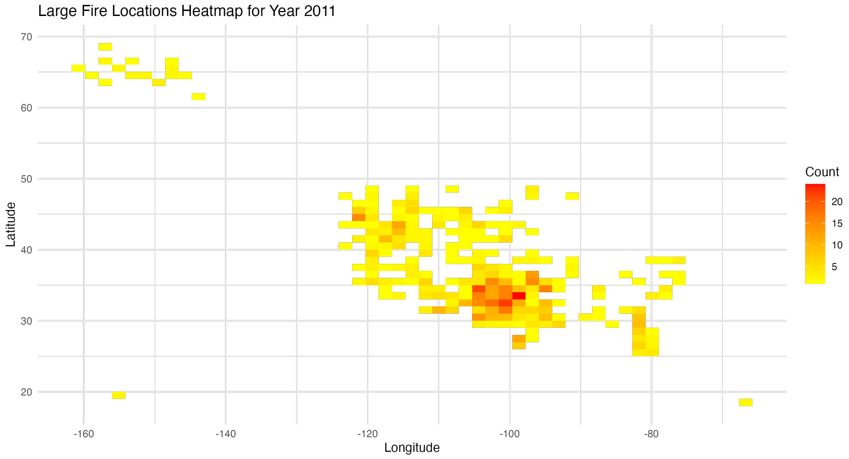

Figure 3.21: Fire Location Heatmap for Large Fires in 2011

For large fires, different from the pattern of small fires, the heatmaps from 2000, 2007, and 2011 (Figures 3.19, 3.20, and 3.21) highlight that central Southern and central Northwestern regions have more frequent large fires. The increase in fire sizes in these regions during these years can be linked to factors such as intense drought conditions, high temperatures, and increased human activity, all of which contribute to the higher likelihood of large fires.

In 2000, large fires in the central U.S. were notably high, with widespread occurrences in states like Arizona and Texas. By 2007 and 2011, these areas continued to show high fire intensity, with a marked increase in large fire size. The role of climate change in exacerbating these conditions cannot be understated, as prolonged droughts, rising temperatures, and shifts in precipitation patterns have been identified as key contributors to the growing intensity of large wildfires in these regions.

3.6 Analysis of the Reporting Agency & Introducing New Variable - Land Type

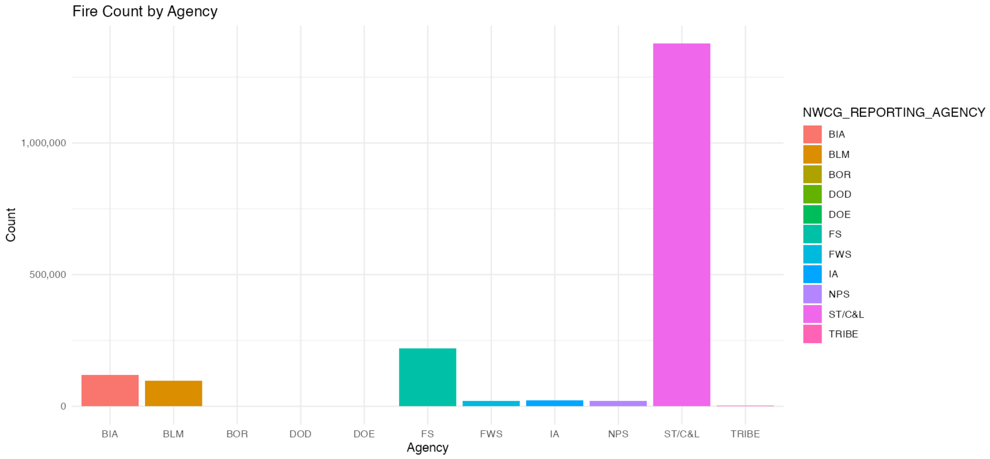

Figure 3.22: Fire Count by Agency

The majority of wildfires are managed by the State, County, or Local Organization (ST/C&L), followed by the U.S. Forest Service (FS), which is responsible for managing vast forested areas. Tribal agencies, Bureau of Indian Affairs (BIA), and the Bureau of Land Management (BLM) also play significant roles in fire management. The high number of wildfires reported by these agencies reflects their extensive land coverage and active fire monitoring and control efforts (Figure 3.22).

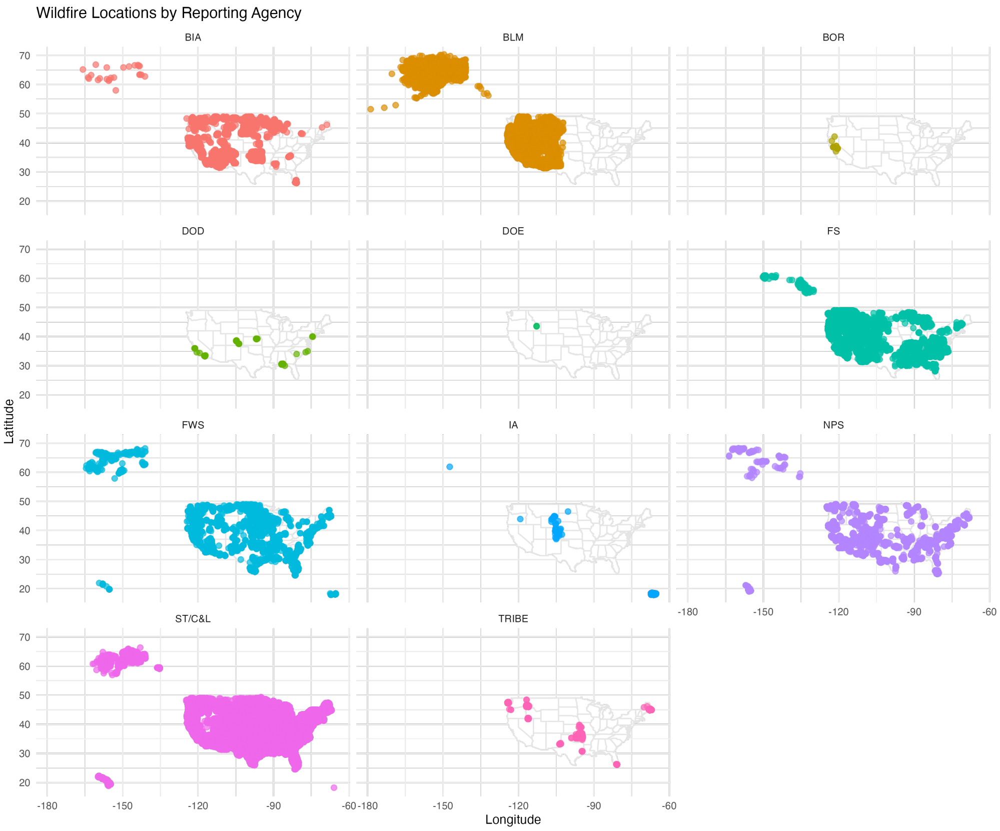

Figure 3.23: Fire Location by Reporting Agency Map

Figure 3.23 provides a spatial representation of fire locations across the United States, grouped by the reporting agency. Each agency is associated with specific types of land and jurisdictions. By analyzing the reporting agency, we can infer the corresponding land type, as each agency typically manages specific types of land. The following categorization highlights the land types associated with each reporting agency:

- **BIA** (Bureau of Indian Affairs): This agency primarily manages Native American lands, which are governed by tribal law. Fires reported by this agency likely occur on reservations and other tribal lands, which can range from forested areas to grasslands. Fires reported by the BIA are primarily located in the western United States and small portions of the Southwest.

- **BLM** (Bureau of Land Management): BLM oversees vast swathes of land, primarily in the western U.S., and the distribution of fire locations largely reflects this extensive coverage.

The fires tend to occur in the rugged terrain of the Rocky Mountains and the surrounding regions. These areas are mostly composed of grasslands, rangelands, and some forested regions, which are typically used for grazing, conservation, and recreation.

- **BOR** (Bureau of Reclamation): BOR is responsible for irrigation projects, water resources, and lands near reservoirs and dams. Fires managed by BOR are typically located in these areas, often related to water conservation or energy production. The distribution shows some fires near the western part of the U.S. around large irrigation systems, with less widespread fire occurrences compared to other agencies due to the nature of BOR's land management.

- **DOD** (Department of Defense): The Department of Defense manages military lands, including bases and training grounds. These sites are scattered, often isolated from civilian land. The fire locations seem to be concentrated around active military facilities, highlighting the potential for high-intensity fires during training activities.

- **DOE** (Department of Energy): DOE manages lands associated with energy production, including nuclear facilities and power plants. Fires here may occur around energy production sites or other restricted areas. The fire locations appear in a few isolated spots, reflecting the restricted nature of DOE-managed lands.

- **FS** (Forest Service): The Forest Service manages national forests and grasslands, often characterized by large, dense forests. Fires reported by the FS are usually found in these protected areas, which are actively managed for conservation, recreation, and biodiversity. From the map, the distribution shows a heavy concentration of fire incidents in these forest regions such as California, Oregon, and Washington, where wildfires are a major concern due to dry conditions and forest ecosystems. The widespread nature of fire locations in these regions indicates a high volume of forest management activities and fire incidents.

- **FWS** (Fish and Wildlife Service): FWS oversees national wildlife refuges and other conservation areas. Fires reported by FWS are typically found in wetlands, wildlife reserves, and other ecologically sensitive areas. The spatial pattern is concentrated in parts of the Southeast U.S., particularly around Florida, where wetland habitats are abundant, and near the coast. These regions tend to experience fires in response to both natural events and fire management practices.

- **IA** (Interagency Organization): This indicates cooperation between multiple agencies. Fires in areas managed by IA could involve shared management responsibilities and may occur in diverse land types. The distribution map shows that fires reported by IA are primarily in central North, such as Wyoming (WY) and Colorado (CO), and Puerto Rico (PR).

- **NPS** (National Park Service): The NPS manages national parks and monuments. Fires in these areas are often linked to natural or recreational activities within these designated conservation and historical sites. The fire distribution is widespread, covering national parks from coast to coast, with high concentrations in the western U.S., where national parks are prevalent.

- **ST/C&L** (State, County, or Local Organization): Fires reported by local and state agencies tend to be clustered in areas with higher human populations. These fires are particularly concentrated in suburban and urban areas, where human activity contributes to the overall fire count. Most of the fires are reported by this agency.

- **TRIBE** (Tribal Organization): Fires reported by tribal organizations likely occur on lands managed by Native American tribes, which could encompass a range of ecosystems such as forests, grasslands, or wetlands. These fires are scattered across various regions, often in the south U.S., where tribal lands are found. The distribution is consistent with the areas managed by BIA but more localized to specific reservations.

Based on the categorization, we can classify the land types into the following categories:

- **Forest**: Includes lands managed by the Forest Service (FS) and National Park Service (NPS), typically covered by forests and managed for conservation, recreation, and biodiversity.

- **Grassland/Range**: Managed primarily by the Bureau of Land Management (BLM), these lands consist of grasslands and rangelands, often used for grazing and land management.

- **Wetlands**: These lands are managed by the Fish and Wildlife Service (FWS) and are characterized by standing water or periodic flooding, often used for wildlife conservation.

- **Urban/Developed**: Managed by state, county, or local organizations (ST/C&L) and Interagency Organizations (IA), these lands are primarily urban or developed areas such as cities, towns, and industrial zones.

- **Water Bodies**: Associated with the Bureau of Reclamation (BOR), these lands are primarily large bodies of water such as reservoirs and lakes, which are used for water conservation or irrigation purposes.

- **Military/Restricted**: Managed by the Department of Defense (DOD) and the Department of Energy (DOE), these lands are primarily used for military training or energy production, often with restricted access.

- **Tribal Land**: Managed by the Bureau of Indian Affairs (BIA) or Tribal Organizations (TRIBE), these lands are located within Native American reservations and governed by tribal law.

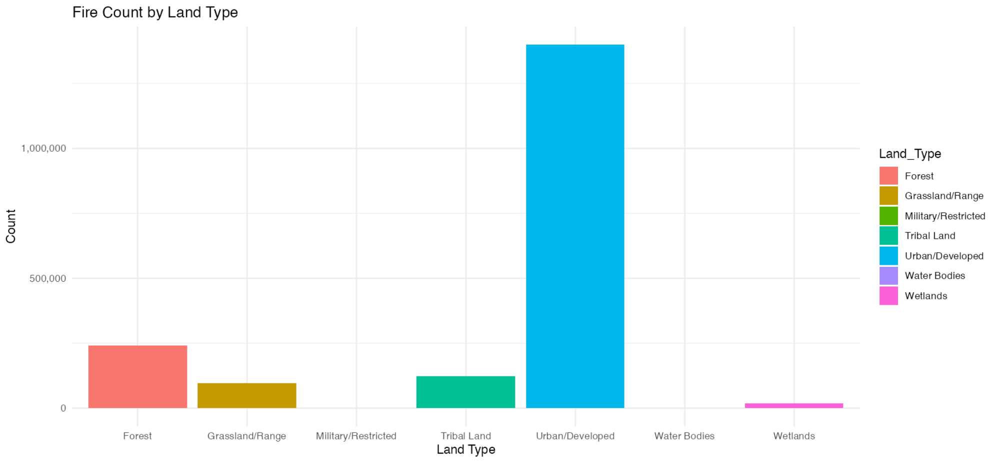

Figure 3.24: Fire Count by Land Type

Figure 3.24 shows the count of wildfires by land type. As seen, the majority of fires occur in urban or developed areas, followed by forested areas and tribal lands. This trend suggests that while fires are widespread, urban areas are particularly prone to frequent smaller fires, likely due to human activity.

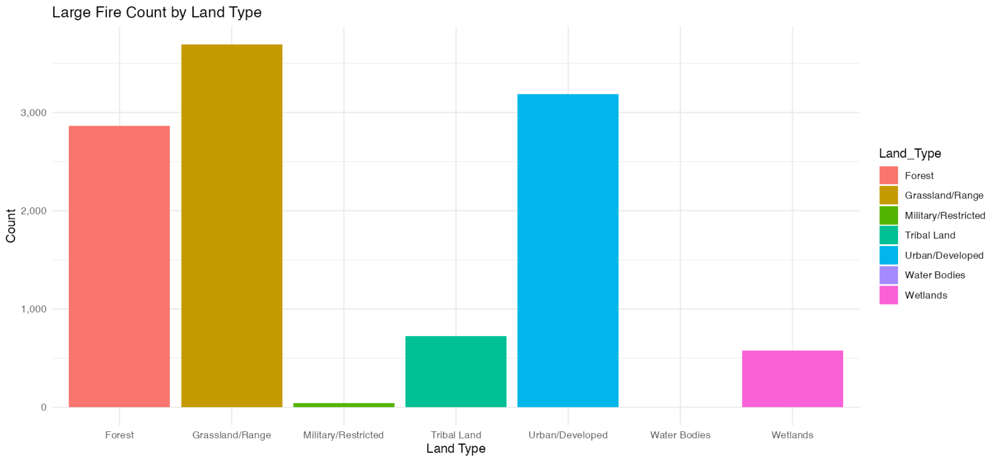

Figure 3.25: Large Fire Count by Land Type

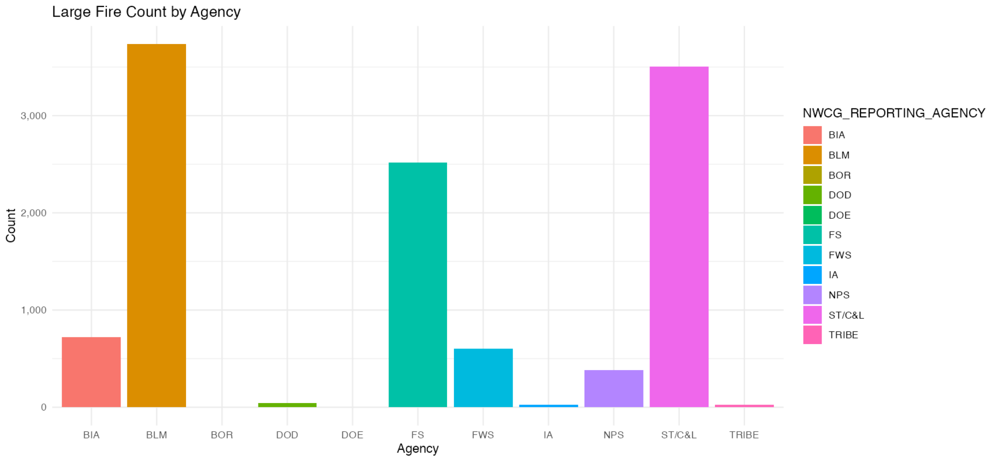

Figure 3.26: Large Fire Count by Agency

As seen in Figure 3.25, large fires are more commonly found in grassland/range areas, followed by urban/developed areas and forested regions. Large fires are most commonly reported by federal agencies such as the FS, ST/C&L, and BLM, as shown in Figure 3.26. These agencies are responsible for managing expansive lands, which are more susceptible to significant fires due to their size and the conditions of the land.

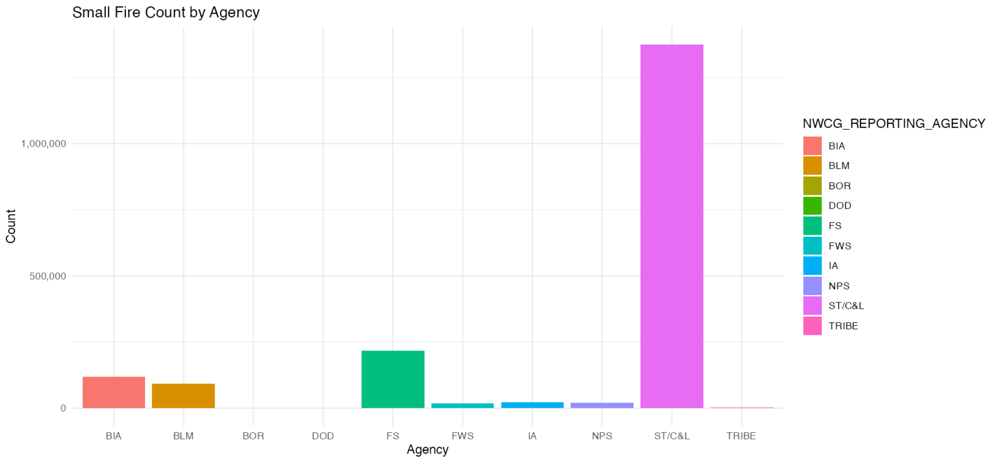

Figure 3.27: Small Fire Count by Agency

In contrast, small fires are predominantly handled by local agencies, as shown in Figure 3.27. These fires often occur in more populated, urbanized areas where human activity is more frequent, leading to a higher occurrence of smaller fires.

3.7 Analysis of the Average Temperature When Fires Occur

In this section, the average temperature across different states and years when fires occur is analyzed. It provides insights into the climate conditions during the occurrence of large and small fires.

3.7.1 Average Temperature by State When Large Fires Occur

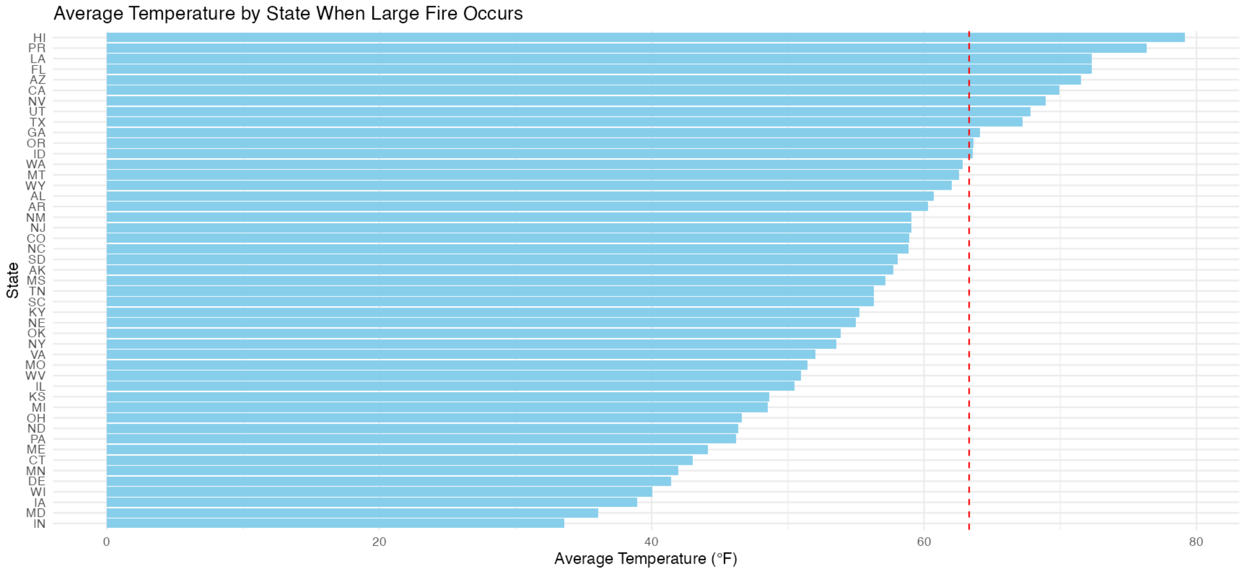

Figure 3.28: Average Temperature by State When Large Fire Occurs

Figure 3.28 illustrates the average temperature in Fahrenheit for each state when large fires occur. As shown, the states with the highest average temperatures when large fires occur are Hawaii (HI), Puerto Rico (PR), and Louisiana (LA). These states are generally characterized by warmer climates, which is consistent with the occurrence of large fires in hotter conditions. The red vertical dashed line in the figure indicates the overall average temperature, which is approximately 63°F. This average is higher for most states with large fires, highlighting the correlation between temperature and the occurrence of larger fires.

3.7.2 Average Temperature by State When Small Fires Occur

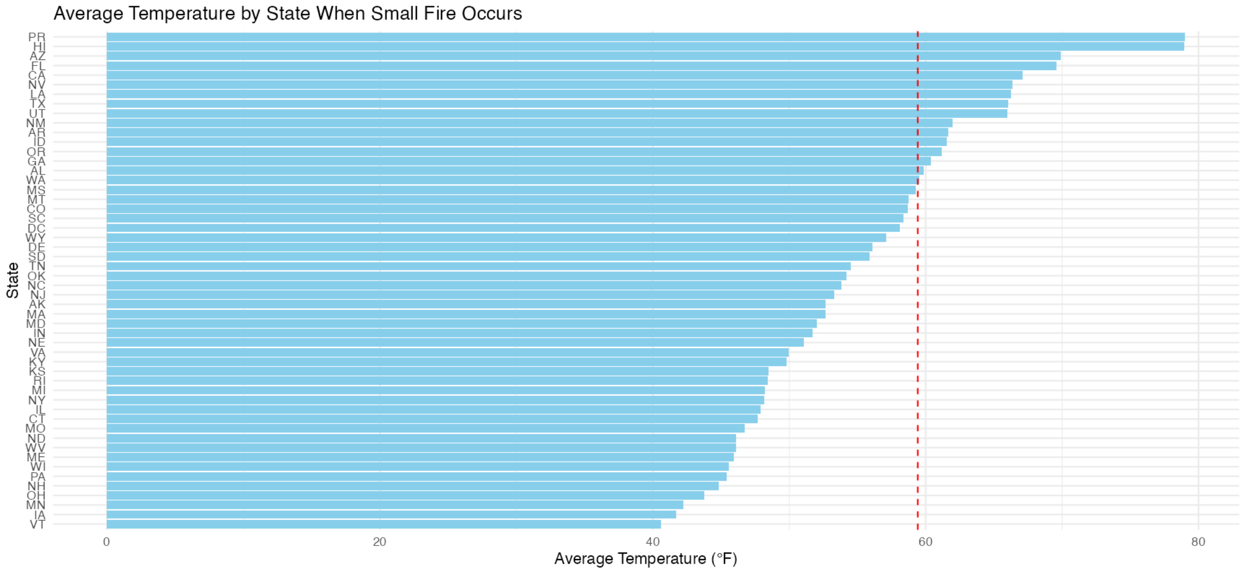

Figure 3.29: Average Temperature by State When Small Fire Occurs

Figure 3.29 shows the average temperature for each state when small fires occur. Similar to large fires, states like Hawaii, Puerto Rico, and Arizona show higher temperatures. However, the overall average temperature for small fires is slightly lower, at about 60°F, as indicated by the red dashed line. This suggests that the overall average temperature when small fires occurs is lower than the overall average temperature when large fires occur.

3.7.3 Average Temperature by Year When Large Fires Occur

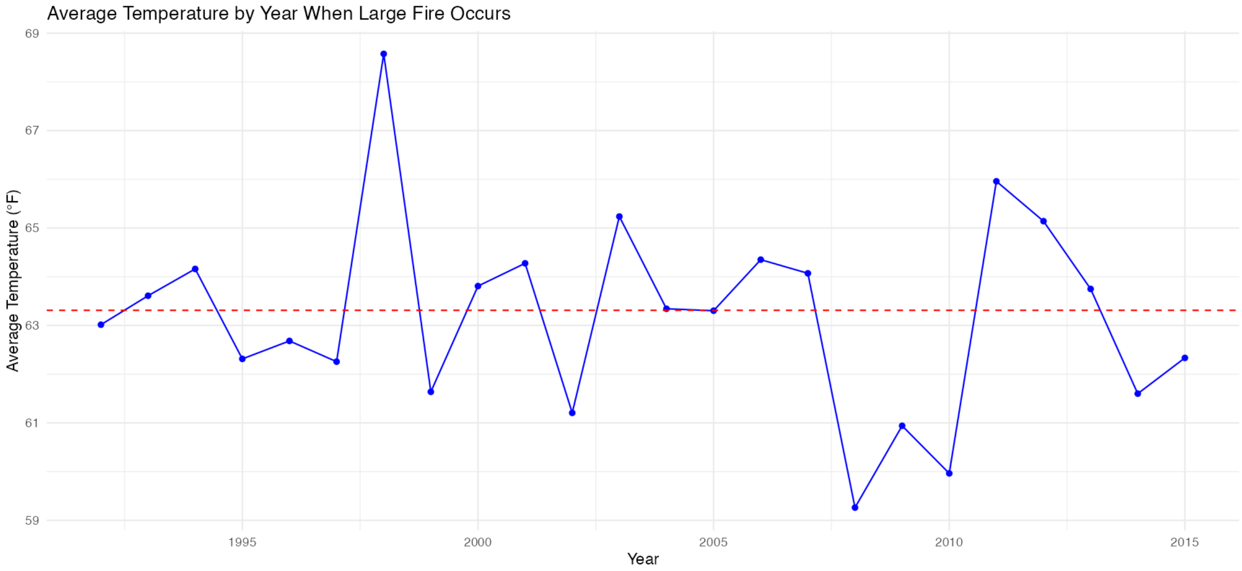

Figure 3.30: Average Temperature by Year When Large Fire Occurs

Figure 3.30 further shows the trend of average temperatures over the years when large fires occur. The average temperature fluctuates significantly, with peaks in some years, such as around 1998 and 2011. Despite these variations, the overall trend line indicates that the average temperature remains around 63°F, as marked by the red dashed line.

3.7.4 Average Temperature by Year When Small Fires Occur

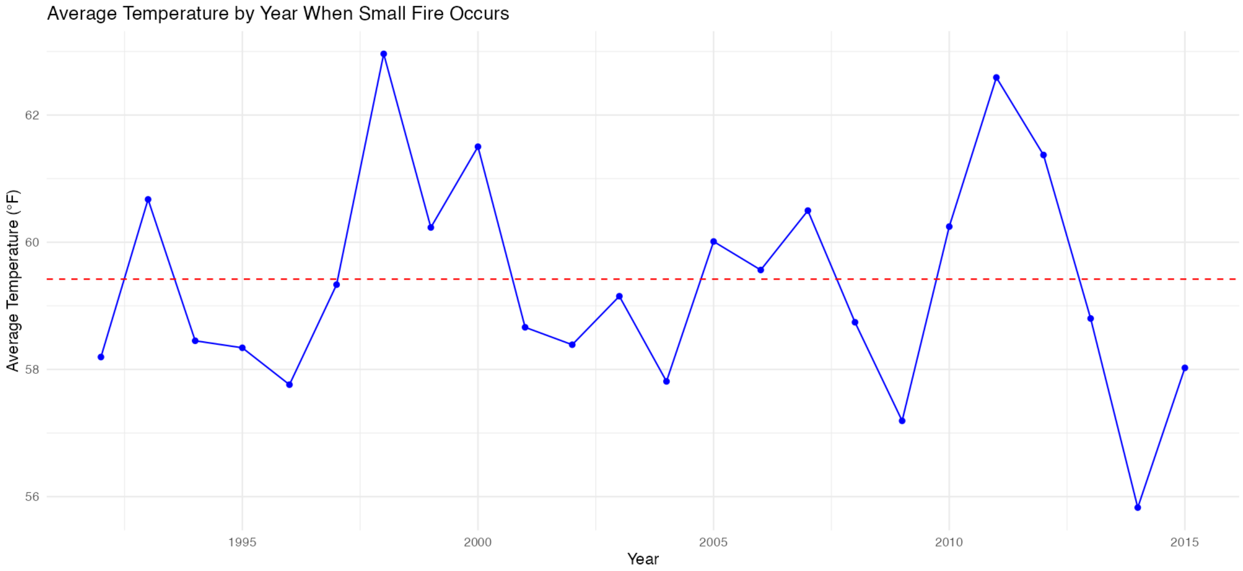

Figure 3.31: Average Temperature by Year When Small Fire Occurs

For small fires, the temperature trend by year also shows noticeable fluctuations (Figure 3.31). The temperature tends to fluctuate more, particularly in recent years.

3.8 Correlation Heatmaps

The correlation heatmaps for both small and large fires provide important insights into the relationships between fire-related numeric variables, including the year of the fire, geographical coordinates (latitude and longitude), average temperature, and fire size. Since the goal is to predict fire size, understanding these relationships is crucial before applying machine learning models such as CatBoost, XGBoost, Random Forest, GLMs, or MLP.

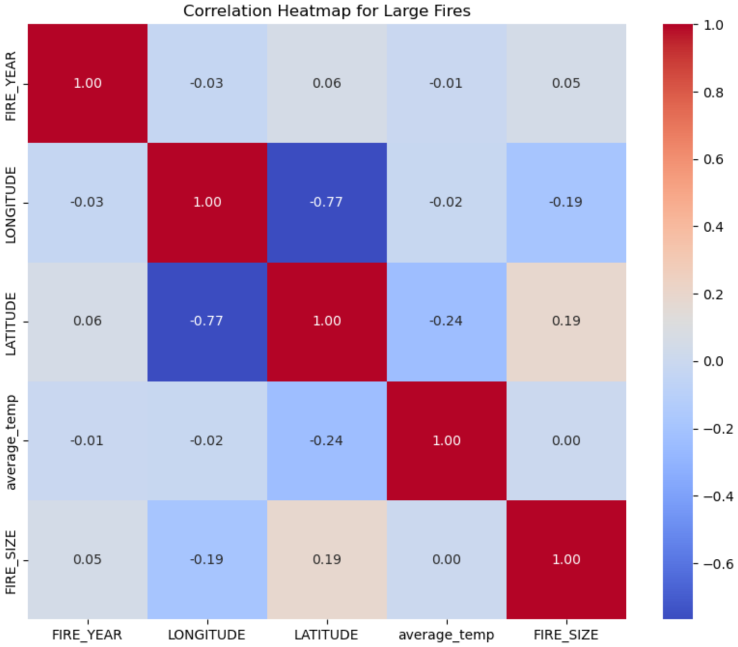

Figure 3.32: Correlation Heatmap for Large Fires

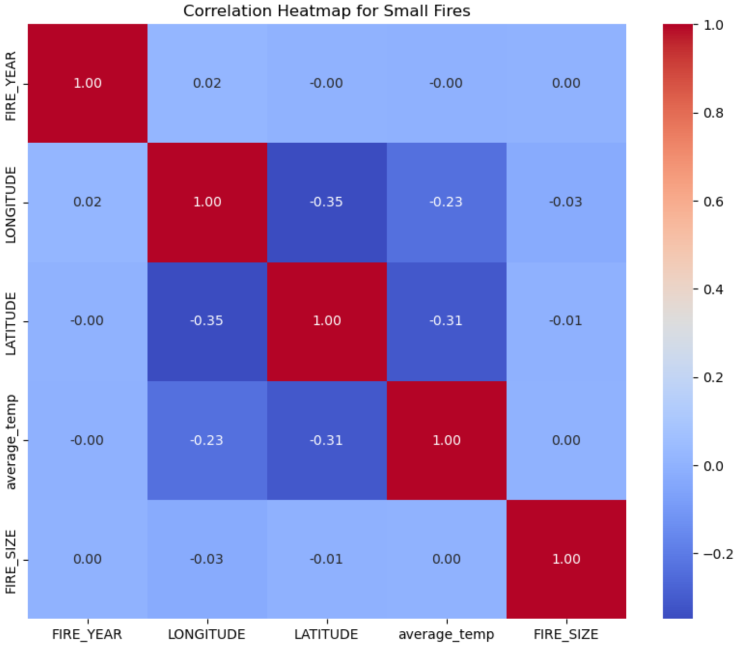

Figure 3.33: Correlation Heatmap for Small Fires

The correlation heatmaps for both large and small fires reveal that fire year and average temperature have weak correlations with fire size, indicating that these variables are not strong predictors

(Figure 3.32, 3.33. The latitude and longitude show moderate correlations with fire size, particularly for small fires, with larger fires occurring more frequently in southern and western regions.

Given these weak to moderate correlations, it is likely that non-linear models, such as Random Forest, XGBoost, and CatBoost, will provide better results for predicting fire size. These models can capture complex, non-linear relationships in the data, making them more suitable for this task compared to linear models.

CHAPTER 4

Methodology

This chapter outlines the machine learning models employed in this study. The primary models used include XGBoost, CatBoost, Random Forest, Generalized Linear Models (GLMs), and Multilayer Perceptron (MLP). These models are applied to different aspects of the study: predicting fire size for small and large fires, and predicting the likelihood of a fire being classified as small or large. The models are trained on a dataset that combines wildfire occurrence data and temperature data across the United States from 1992 to 2015, with features such as fire size, location, cause, and temperature. The models are evaluated using Root Mean Squared Error (RMSE) as the primary performance metric. To address the skewness and heteroscedasticity in the fire size data, the log transformation was applied to the target variable to improve model performance and reduce RMSE.

The fire dataset is obtained from Kaggle [15], and the temperature data is sourced from NOAA's National Centers for Environmental Information (NCEI) [9]. Based on the reporting agency feature in the fire data, we created a new variable, land type. The final dataset, after preprocessing, includes the following variables: ID, fire year, state, cause, fire size, land type, longitude, latitude, and average temperature.

4.1 Log Transformation

The log transformation is commonly used to reduce skewness and stabilize variance, particularly in datasets with heavy tails or where the target variable exhibits heteroscedasticity. It is particularly effective when the data spans multiple orders of magnitude or when extreme values (outliers) may influence the results. The log transformation compresses the range of the data, making it more suitable for modeling with methods that assume normality or constant variance.

Mathematically, the log transformation is expressed as:

$$y_{\text{log}} = \log(y)$$

In this formula, y represents the original target variable (fire size), and y_{log} is the transformed target variable. The natural logarithm is typically used, as it stabilizes the variance and reduces the influence of large values by compressing the scale. In this study, the log transformation was applied to the fire size data to address the skewed distribution and heteroscedasticity of the target variable. The transformation was expected to improve the model's ability to predict fire sizes by stabilizing variance. It is noticeble to mention that other transformation methods are also tried like the Box-Cox Transformation where it fails to meet its assumption.

After applying the log transformation, the data is used to train the models. Once the models make predictions, the inverse log transformation is applied to bring the predictions back to the original scale of fire size:

$$y = \exp(y_{\text{log}})$$

This step ensures that the model's predictions are in the same units as the original target variable. By applying the log transformation, the models are better able to handle skewed data and non-constant variance, leading to improved predictive performance for large fire size predictions. However, for small fire predictions, the log transformation did not produce as significant an improvement as expected, and the distribution still shows some skewness.

The following figures illustrate the effect of the log transformation on the fire size distribution for both small and large fire data. Although the transformation does not fully normalize the distribution, it still makes the data more suitable for modeling compared to the original, especially for large fires.

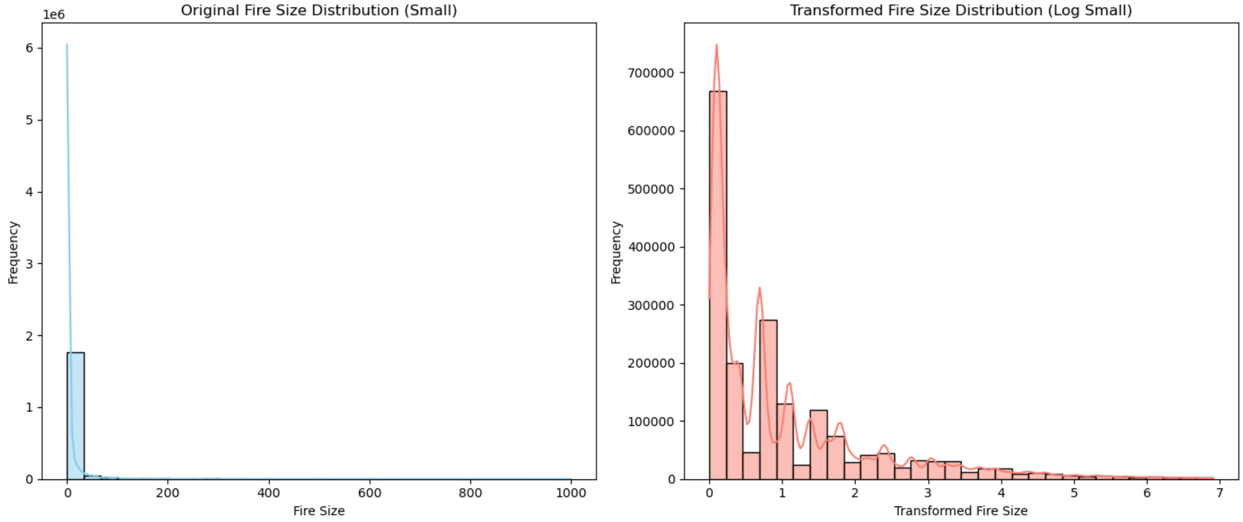

Figure 4.1: Left: Original fire size distribution (Small); Right: Transformed fire size distribution after Log Transformation (Small).

As seen in Figure 4.1, the original distribution of small fire sizes is heavily skewed to the right, with a long tail of large values. After applying the log transformation, the distribution becomes more symmetric, reducing the skewness and making it closer to a normal distribution. This transformation helps stabilize the variance, which is crucial for improving model performance, though it did not lead to as substantial a reduction in RMSE for small fire predictions.

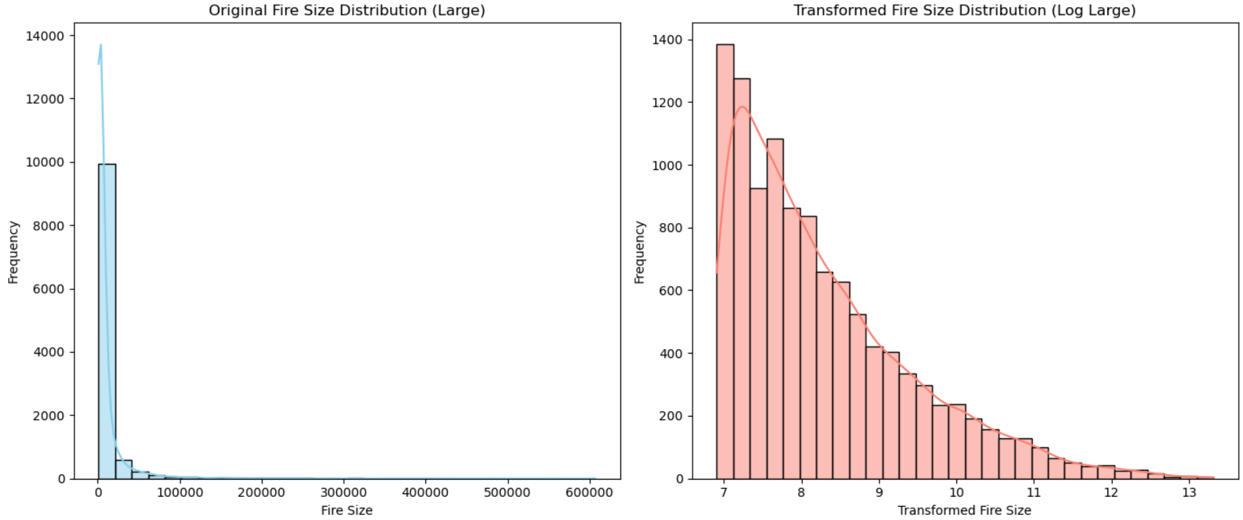

Figure 4.2: Left: Original fire size distribution (Large); Right: Transformed fire size distribution after Log Transformation (Large).

Similarly, for large fire data, the log transformation also results in a more symmetric distribution. As seen in Figure 4.2, the original distribution is highly skewed with large values, which could cause problems during modeling due to the presence of extreme values. After applying the log transformation, the distribution becomes more uniform, making the data more suitable for predictive modeling. The transformed data allows the models to learn more effectively from the patterns in the data. Although the log transformation does not fully resolve the skewness, it helps models perform better for large fires compared to the original untransformed data.

4.2 Regression Metric: Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is a widely used metric for evaluating the performance of regression models, providing a measure of how well the model's predictions match the actual values. RMSE calculates the square root of the average squared differences between predicted and actual values, thus giving more weight to larger errors. This makes RMSE particularly useful in situations where large deviations are more problematic and should be penalized more heavily. The formula for RMSE is given by:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where y_i represents the actual value, $\hat{y}_i$ represents the predicted value, and n is the total number of observations. By squaring the residuals (the differences between the predicted and actual values), RMSE amplifies the effect of large errors, making it highly sensitive to outliers in the data.

In this study, RMSE was chosen as the primary evaluation metric for all machine learning models used to predict fire size and classify fires as small or large. Since fire size can exhibit high skewness and large outliers, RMSE is well-suited for assessing model performance in this context. A lower RMSE indicates a better model fit, as it suggests smaller errors in prediction and greater accuracy.

One of the key reasons RMSE was selected is its ability to highlight the impact of large prediction errors. Given that extreme values can significantly influence model training, RMSE helps us to gauge how well the models handle such outliers. Furthermore, RMSE was particularly useful in evaluating the effectiveness of the log transformation, which was applied to the fire size data to reduce skewness and stabilize variance. After applying this transformation, the model's performance was evaluated using RMSE to determine whether the transformation led to more accurate predictions of fire size by improving model stability and reducing error.

4.3 Classification Metrics

For the classification model in this study, multiple metrics are used to provide a comprehensive assessment of how well the model can distinguish between different classes, in this case, small and large fires. The classification metrics discussed here include classification accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, and Matthews Correlation Coefficient (MCC).

Classification accuracy measures the proportion of correct predictions made by the model. It is calculated as the ratio of correct predictions (both true positives and true negatives) to the total number of predictions. The formula for classification accuracy is:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Where:

- TP = True Positives (correctly predicted large fires),
- TN = True Negatives (correctly predicted small fires),
- FP = False Positives (incorrectly predicted large fires),
- FN = False Negatives (incorrectly predicted small fires).

Although accuracy is intuitive, it may not be informative when there is class imbalance, such as when small fires far outnumber large fires. Therefore, additional metrics are necessary to provide a deeper understanding of model performance.

Precision measures the proportion of positive predictions (large fires) that are actually correct. It is especially important when the cost of false positives is high, such as when resources are misallocated to fighting fires that are incorrectly predicted as large. The formula for precision is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Where:

- TP = True Positives (correctly predicted large fires),
- FP = False Positives (incorrectly predicted large fires).

Recall, also known as sensitivity or the true positive rate, measures the proportion of actual positive instances (large fires) that were correctly identified by the model. It is critical when the cost of missing positive cases (false negatives) is high, as failing to identify a large fire could lead to significant consequences. The formula for recall is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Where:

- TP = True Positives (correctly predicted large fires),
- FN = False Negatives (incorrectly predicted small fires).

The F1-score is the harmonic mean of precision and recall. It provides a balanced evaluation by considering both the ability to correctly identify large fires (precision) and the ability to capture as many large fires as possible (recall). The F1-score is particularly useful in imbalanced datasets where accuracy may be misleading. It is calculated by multiplying precision and recall and then dividing by their sum:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score ranges from 0 to 1, with 1 indicating perfect precision and recall, and 0 indicating poor performance.

ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) is a performance measurement for classification problems that evaluates the trade-off between the true positive rate (recall) and the false positive rate. It plots the true positive rate against the false positive rate at various threshold values and calculates the area under this curve. The formula for AUC is:

$$\text{AUC} = \int_{-\infty}^{\infty} \text{True Positive Rate}\, d(\text{False Positive Rate})$$

AUC values range from 0 to 1, with values closer to 1 indicating a better performing model. An AUC of 0.5 suggests that the model is performing no better than random chance.

The confusion matrix provides a summary of the model's predictions, showing the number of true positives, false positives, true negatives, and false negatives. It helps in understanding where the model is making errors. The confusion matrix is a 2×2 table for binary classification, as shown below:

$$\begin{bmatrix} \text{TP} & \text{FP} \\ \text{FN} & \text{TN} \end{bmatrix}$$

Where TP represents true positives, FP represents false positives, FN represents false negatives, and TN represents true negatives.

Finally, the Matthews Correlation Coefficient (MCC) is a more balanced measure of classification performance, especially in imbalanced datasets. It considers all four categories in the confusion matrix (true positives, false positives, false negatives, and true negatives) and provides a single value that reflects the overall performance of the model. The MCC is calculated as:

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

MCC ranges from -1 to 1, where 1 indicates perfect classification, -1 indicates total disagreement, and 0 indicates random prediction.

These classification metrics, including accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, and MCC, provide a comprehensive evaluation of a classification model's performance. They are especially important when the dataset is imbalanced, ensuring that the model's ability to distinguish between small and large fires is adequately assessed across various thresholds and error types.

4.4 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an efficient and scalable implementation of gradient boosting, widely used for both regression and classification tasks. It builds an ensemble of weak learners, typically decision trees, where each successive tree attempts to correct the errors of the previous trees by learning from the residuals (errors) left by earlier iterations. This iterative process improves the predictive performance of the model by gradually refining its predictions. XGBoost optimizes the model's performance by employing both first- and second-order derivatives of the loss function, which accelerates the learning process and enhances model accuracy.

The objective function in XGBoost is formulated as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where $l(y_i, \hat{y}_i)$ represents the loss function, with y_i being the true value and $\hat{y}_i$ the predicted value, and $\Omega(f_k)$ is a regularization term applied to each tree to control model complexity and avoid overfitting. The first summation term minimizes prediction errors, while the second summation term prevents the model from becoming too complex by penalizing large trees, ensuring that the model generalizes well to unseen data.

In this study, XGBoost is applied for predicting both the size of small fires and large fires. This model is particularly suited for small fire predictions due to the large number of small fire instances in the dataset, providing the model with enough data to effectively capture patterns and trends.

The features used for model training include the fire year, state, cause, land type, geographical coordinates (longitude and latitude), and average temperature for the region. These features are incorporated as inputs to predict fire size, which serves as the target variable for the regression task.

XGBoost's ability to handle missing data, process categorical features via one-hot encoding, and manage highly unstructured data makes it ideal for predicting wildfire characteristics. After training, the model predicts the fire size for the test set, and the Root Mean Squared Error (RMSE) is calculated to assess its performance. A lower RMSE indicates a better model fit.

In addition to the regression task, XGBoost is also applied to a classification task where the goal is to predict the likelihood of a fire being classified as small or large. For this classification model, the target variable is binary, indicating whether a fire is classified as small or large. Given the imbalanced nature of the dataset (more small fires than large fires), class weights are introduced to balance the influence of each class during model training. XGBoost's objective function for classification is set to binary logistic regression, making it suitable for binary classification tasks.

The features used for the classification task are the same as those used in the regression task, including fire year, state, cause, land type, longitude, latitude, and temperature. After preparing the data through one-hot encoding of categorical variables, the dataset is split into training and testing sets. The model is trained using the training dataset, and predictions are made on the test set. The model's performance is then evaluated using classification accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, and Matthews Correlation Coefficient (MCC).

XGBoost is particularly beneficial in wildfire prediction due to its capacity to handle complex relationships and its ability to adapt to imbalanced datasets, making it a suitable choice for both regression and classification tasks in this study.

4.5 Categorical Boosting (CatBoost)

Categorical Boosting (CatBoost) is another gradient boosting algorithm that is specifically designed to handle categorical features efficiently. Unlike traditional gradient boosting methods like

XGBoost, which require manual preprocessing of categorical features (e.g., one-hot encoding or label encoding), CatBoost automates the handling of categorical variables. It uses a specialized algorithm based on ordered target statistics to encode categorical features. This method involves the computation of target statistics for categorical variables in a way that reduces overfitting, as it combines permutations of the categorical features during training. In particular, each categorical feature is encoded by considering the statistics of the target variable, such as the mean of the target for each category, while preserving the order of the categories. This technique ensures that the model does not overfit the categorical features, as the encoding depends on the distribution of the target variable within each category rather than on arbitrary assignments of numerical values.

The general objective function for CatBoost can be written as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) + \sum_{c=1}^{C} \mathcal{R}(c)$$

In this formulation, $l(y_i, \hat{y}_i)$ represents the loss function, which measures the discrepancy between the true value y_i and the predicted value $\hat{y}_i$. The loss function is crucial for quantifying how well the model is performing, and it guides the optimization process by providing feedback on the error between the predicted and actual values. For regression tasks, the loss function typically used is the Root Mean Squared Error (RMSE), which penalizes larger errors more significantly, encouraging the model to make more accurate predictions. In classification tasks, a log-loss or binary cross-entropy function is often used, as it measures the error between the predicted probability and the actual class labels. The loss function is an essential component of the model's objective, as it defines how the model learns to minimize prediction errors during training.

Additionally, the term $\Omega(f_k)$ represents the regularization applied to each tree to control its complexity and prevent overfitting. This regularization term penalizes overly complex trees that may fit noise in the data rather than capturing the underlying patterns. The final term, $\mathcal{R}(c)$, specifically applies to the categorical features and ensures that their encoding remains smooth. This term regularizes the encoded values of the categorical features, preventing the model from overfitting by excessively relying on any one category. Together, these components of the objective function allow CatBoost to build an effective and efficient model that can handle complex categorical data without overfitting.

In this study, CatBoost is used for both small and large fire predictions. The algorithm is particularly effective for handling complex categorical data involved in predicting wildfire dynamics. Features such as the fire's cause, the state in which it occurred, and the land type are categorical in nature, and CatBoost processes these variables with minimal manual preprocessing. CatBoost's ability to perform well with default hyperparameters is a significant advantage, especially when there is limited time or resources for model tuning. This capability is particularly useful in the context of wildfire prediction, where the features may be varied and complex.

For small fire predictions, CatBoost provides high prediction accuracy due to the large and diverse dataset available for training. The large number of small fires offers a rich set of patterns that the model can learn from, allowing it to make accurate predictions about fire size. On the other hand, when predicting large fires, the dataset is more sparse, and although CatBoost still provides valuable insights, the predictions are less accurate due to the limited data available for training.

The performance of CatBoost is assessed in a manner similar to that of XGBoost. For the regression task, the model's performance is evaluated by comparing the predicted fire sizes to the actual fire sizes, with a lower RMSE indicating a better model fit. Cross-validation is used to assess the robustness of the model and to prevent overfitting, especially when training data is limited.

In addition to the regression task, CatBoost is applied to a classification task, where the goal is to predict whether a fire is small or large. The classification task involves a binary target variable, with '1' representing large fires and '0' representing small fires (using a threshold of 1000 acres). CatBoost's handling of categorical variables is particularly useful here as well, as it can process the categorical features efficiently. The model is trained using the same features as in the regression task, with the target variable now being binary. For this classification task, the performance of CatBoost is evaluated using the same metrics as the XGBoost.

CatBoost's ability to directly and efficiently handle categorical features makes it an excellent choice for modeling the complex relationships in wildfire data, particularly when working with mixed data types. By leveraging its powerful handling of categorical variables and default hyperparameters, CatBoost offers a robust and efficient solution for both regression and classification tasks in wildfire prediction.

4.6 Random Forest

Random Forest is an ensemble learning algorithm that consists of multiple decision trees, where each tree is trained on a random subset of the data using a technique called bootstrap sampling. The final prediction is obtained by aggregating the predictions from all the individual trees, which helps to reduce overfitting and enhances the model's generalization ability. The use of multiple trees significantly reduces the variance of the model's predictions, making it robust even when the dataset contains noise or highly complex relationships. Random Forest combines two key techniques—bootstrapped sampling and random feature selection—to build diverse and highly accurate trees.

The process of training a Random Forest model can be described as follows: First, from the original dataset, multiple bootstrap samples are drawn. A bootstrap sample is generated by randomly selecting samples from the dataset with replacement, meaning that some data points may be selected multiple times while others may not be selected at all. For each decision tree, a random subset of features is also selected at each split. This is known as random feature selection, and it ensures that each tree in the forest is built using a different combination of features. This randomness in both the data and the features contributes to the diversity of the trees, reducing the likelihood of overfitting and improving the model's performance.

The prediction for a given input x_i in a Random Forest is calculated as the average of the predictions from all the individual trees. For regression tasks, this aggregation is performed by averaging the predicted values from each tree:

$$\hat{y}_i = \frac{1}{T}\sum_{t=1}^{T} f_t(x_i)$$

where T is the number of trees in the forest, and $f_t(x_i)$ is the prediction of the t-th tree for the input x_i. For classification tasks, the prediction is typically made by taking a majority vote from all trees in the forest, with each tree casting a vote for the predicted class.

One of the key advantages of Random Forest is that by aggregating the predictions from multiple decision trees, it significantly reduces the variance of the model. This means that the model is less likely to overfit the data, even when the individual trees may overfit on their own. Additionally, because each tree is trained on a different random subset of the data and features, Random Forest is able to capture different patterns and interactions within the data, resulting in a more robust and generalized model.

For small fire predictions, where the dataset is large and diverse, Random Forest performs well by capturing the complex, nonlinear relationships between various features, such as land type, cause, temperature, and geographic location. The diversity of the trees allows the model to identify intricate patterns in the data, leading to accurate predictions. However, for large fires, where the dataset is sparse, the model may experience higher variance in its predictions. Despite this, Random Forest still provides more reliable estimates compared to individual decision trees, as the ensemble of trees helps to smooth out the errors caused by the limited data for large fires.

The performance of Random Forest is evaluated using standard metrics for both regression and classification tasks. For the regression task, the model's performance is typically assessed using Root Mean Squared Error (RMSE), which measures the average squared difference between the predicted fire sizes and the actual fire sizes. Lower RMSE values indicate better model fit. Cross-validation is employed to ensure that the model generalizes well to unseen data and is not overfitting to the training set.

For the classification task, Random Forest is used to predict whether a fire is small or large, based on a binary target variable (with '1' representing large fires and '0' representing small fires). In this case, class imbalance may be present, as small fires are more common than large fires. To address this, Random Forest can be configured to apply class weights to balance the influence of each class, ensuring that the model does not disproportionately favor the majority class. The model is trained using the same features as in the regression task, and performance is evaluated using the metrics mentioned in section 4.3.

One of the strengths of Random Forest is its ability to handle high-dimensional and mixed-type data. It can automatically handle both numerical and categorical features, and its inherent ability to reduce overfitting through bootstrapping and random feature selection makes it a powerful tool for complex prediction tasks such as wildfire prediction. By aggregating multiple decision trees, Random Forest captures a wide variety of relationships and interactions in the data, leading to more accurate and reliable predictions for both regression and classification tasks.

4.7 Generalized Linear Models (GLMs)

Generalized Linear Models (GLMs) are an extension of traditional linear regression, allowing the dependent variable to follow any distribution from the exponential family, such as Poisson, Binomial, or Gaussian distributions. GLMs use a link function to establish the relationship between the dependent variable and the predictors. For this study, the GLM was applied using a Gaussian family with an identity link function, which assumes a linear relationship between the dependent variable (fire size) and the independent variables (such as temperature, cause, and location). The general form of a GLM is expressed as:

$$g(E[Y|X]) = X\beta$$

where g is the link function, $E[Y|X]$ is the expected value of the dependent variable given the independent variables, X is the matrix of predictors, and β represents the coefficients of the predictors. In the case of fire size prediction, the GLM with a Gaussian family assumes that the fire size follows a normal distribution, and the link function applied was the identity function, meaning the fire size is modeled directly as a linear combination of the predictors.

In this study, GLMs were employed to model both small and large fire sizes. GLMs offer the advantage of providing clear, interpretable relationships between predictors and the dependent variable. However, despite applying several transformations to the dependent variable the assumptions underlying the GLM were not met. These assumptions include linearity, homoscedasticity, and the normality of residuals.

The primary assumption of GLMs is that the relationship between the dependent variable and the predictors is linear in the transformed scale (after applying the link function). However, even after log-transforming and Box-Cox transforming the dependent variable, the relationship between fire size and the predictors remained non-linear, which violated the linearity assumption. Furthermore, GLMs assume homoscedasticity, meaning the variance of the residuals should be constant across all levels of the fitted values. Despite applying transformations, the residuals showed increasing variance as the fitted values increased, suggesting the presence of heteroscedasticity. This pattern indicates that the variance of the residuals was not constant, which violates the assumption of homoscedasticity. Additionally, the normality assumption of GLMs was not satisfied. The residuals did not follow a normal distribution, which is important for making valid inferences about the model coefficients and for performing hypothesis testing.

Although the GLM was used successfully in other studies, the assumptions of linearity, homoscedasticity, and normality of residuals were not satisfied in this particular case. Despite efforts to address these issues through various transformations of the dependent variable, the GLM was ultimately not suitable for predicting fire size in this dataset. As a result, a different modeling approach was explored to address these limitations, which is discussed in the **Results** chapter. The **Results** chapter will provide a more detailed comparison of alternative models that may better fit the characteristics of the wildfire data and provide more accurate predictions.

4.8 Multi-layer Perceptron (MLP)

A Multi-layer Perceptron (MLP) is a type of artificial neural network that consists of multiple layers of neurons. It is a supervised learning algorithm that can learn complex, non-linear relationships between input features and the target variable. The architecture of an MLP includes three primary layers: the input layer, one or more hidden layers, and the output layer. Each neuron in a given layer is connected to all neurons in the next layer, and each of these connections has a weight that is adjusted during training. The hidden layers apply non-linear transformations to the input data, which enables the MLP to capture intricate patterns in the dataset.

In this study, the MLP model is used to predict the fire size. The target variable is the fire size, and the features used for prediction include fire year, state, cause, land type, geographical coordinates (longitude and latitude), and average temperature. These features are input into the model's input layer, and the transformed data is passed through one or more hidden layers, where non-linear transformations occur, ultimately producing the predicted fire size in the output layer.

The architecture of the MLP starts with the input layer, which receives the feature values for each observation. These inputs are then passed through one or more hidden layers. In each hidden layer, the neurons perform a weighted sum of the input features, add a bias term, and then apply an activation function to introduce non-linearity. The output from each neuron in the hidden layers is computed as:

$$h_k = f\left(\sum_{j=1}^{p} w_{kj} x_j + b_k\right)$$

where h_k is the output of the k-th neuron in the hidden layer, w_{kj} is the weight of the connection between the j-th input feature and the k-th hidden neuron, x_j is the j-th input feature, and b_k is the bias term for the k-th neuron. The activation function f is applied to the weighted sum of the inputs, introducing non-linearity to the model. In this study, the ReLU (Rectified Linear Unit) activation function is used for the hidden layers, which is defined as:

$$f(x) = \max(0, x)$$

ReLU is widely used because it is computationally efficient and helps mitigate the vanishing gradient problem, allowing the model to learn faster and perform better. If the input to a neuron is positive, it is passed through unchanged; if the input is negative, it is set to zero. This property allows MLP models to capture non-linear relationships between the input features and the target variable, making it suitable for tasks such as predicting fire size.

The output layer of the MLP produces the final prediction. In this case, the prediction is the estimated fire size, which is computed as a weighted sum of the activations from the last hidden layer:

$$\hat{y} = \sum_{k=1}^{n} w_k h_k + b$$

where $\hat{y}$ is the predicted fire size, w_k is the weight associated with the k-th neuron in the final hidden layer, h_k is the output from the k-th neuron in the last hidden layer, and b is the bias term for the output neuron. The model is trained to minimize the error between the predicted fire size $\hat{y}$ and the actual fire size y.

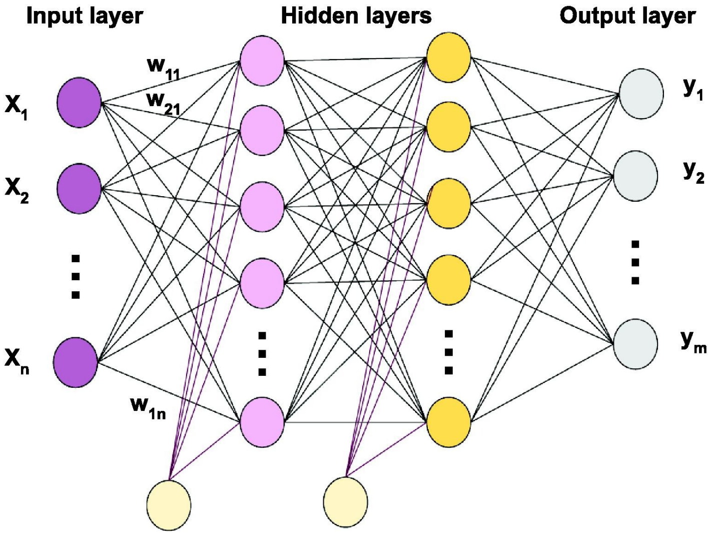

Figure 4.3: Representation of the MLP architecture with input, hidden, and output layers (Source: Chan et al., 2023 [4]).

The training process of the MLP involves the use of backpropagation and gradient descent to update the weights and biases. Backpropagation computes the gradients of the loss function with respect to each parameter in the model, and these gradients are used to adjust the weights and minimize the error. The Mean Squared Error (MSE) is used as the loss function for training, and it is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the actual fire size and $\hat{y}_i$ is the predicted fire size for the i-th observation. MSE is used because it provides a smooth and differentiable function that can be easily optimized using gradient descent. A lower MSE indicates that the model's predictions are closer to the true fire sizes.

The model's weights are updated using the following rule in gradient descent:

$$w_{kj}^{(t+1)} = w_{kj}^{(t)} - \eta \frac{\partial \mathcal{L}}{\partial w_{kj}}$$

where $w_{kj}^{(t+1)}$ is the updated weight, $w_{kj}^{(t)}$ is the current weight, η is the learning rate, and $\frac{\partial \mathcal{L}}{\partial w_{kj}}$ is the gradient of the loss function with respect to the weight w_{kj}. The learning rate η controls the step size for updating the weights during training.

After training, the model's performance is evaluated using the Root Mean Squared Error (RMSE), which is derived from MSE. RMSE is used to assess the model's performance by measuring the difference between the predicted and actual fire sizes. RMSE is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

A lower RMSE indicates better model accuracy, as it shows that the model's predictions are closer to the actual fire sizes.

CHAPTER 5

Results and Discussion

This chapter presents the results of the study, focusing on the performance of the machine learning models used to predict small and large fire sizes, as well as the classification of fire size. The analysis includes model evaluation metrics, specifically the Root Mean Squared Error (RMSE), before and after applying the log transformation to the target variable. The findings highlight the effectiveness of each model and the impact of the transformation on prediction accuracy.

5.1 Small Fire Size Prediction

This section presents the results of predicting small fire sizes using five machine learning models: XGBoost, CatBoost, Random Forest, Generalized Linear Model (GLM), and Multi-layer Perceptron (MLP). We evaluate the models' performance based on their Root Mean Squared Error (RMSE) both before and after applying the log transformation to the target variable. The transformation aims to improve normality and stabilize variance, potentially enhancing the models' prediction accuracy.

5.1.1 RMSE Results for Small Fire Size Prediction

The RMSE values for each model before and after the log transformation are summarized in Table 5.1. The RMSE metric is used to evaluate the difference between the predicted fire size and the actual fire size, with lower values indicating better model performance.

Model	RMSE (Before Transformation)	RMSE (After Log Transformation)
XGBoost	51.46	53.82
CatBoost	51.51	53.73
Random Forest	52.42	52.65
GLM	53.13	54.36
MLP	52.70	54.38

Table 5.1: RMSE values for predicting small fire size before and after applying log transformation.

5.1.2 Analysis and Comparison of Results

Prior to transformation, RMSE values ranged from 51.46 to 53.13 across all models. After the log transformation, RMSE values slightly increased, suggesting that the transformation did not significantly improve prediction accuracy for small fires.

For XGBoost, the model performed well initially with an RMSE of 51.46. However, after applying the log transformation, the RMSE increased to 53.82, indicating that the transformation did not enhance its predictive accuracy for small fire sizes. This is not surprising since XGBoost, as a gradient boosting model, is already be well-equipped to handle skewed and heteroscedastic data without requiring a transformation. The model is designed to learn from the residuals of previous trees, and this inherent flexibility may make it less sensitive to the kind of variance stabilization that a log transformation typically provides.

CatBoost showed similar performance since it may have been capable of capturing the underlying patterns of the data without needing the transformation. The model's robustness in dealing with different types of data may reduce the necessity for transformation-based improvements.

For Random Forest, the model performed adequately before the transformation with an RMSE of 52.42. After applying the log transformation, the RMSE is similar. The reason is that Random Forest models, due to their ensemble nature, are less sensitive to specific data distributions and can handle non-normality relatively well. The transformation did not lead to a substantial improvement likely because the model's underlying decision trees were already robust to the skewed data.

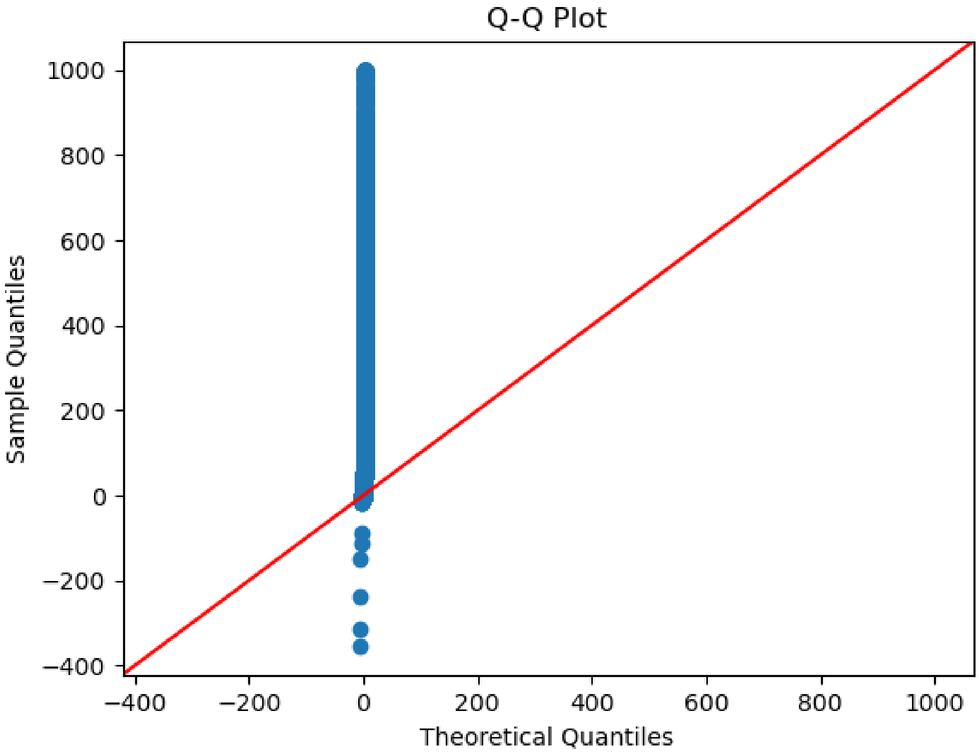

Figure 5.1: GLM with Log Transformation Q-Q Plot

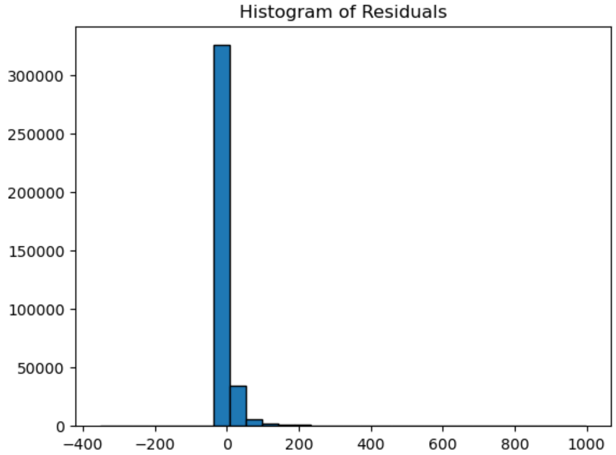

Figure 5.2: GLM with Log Transformation Residual Histogram

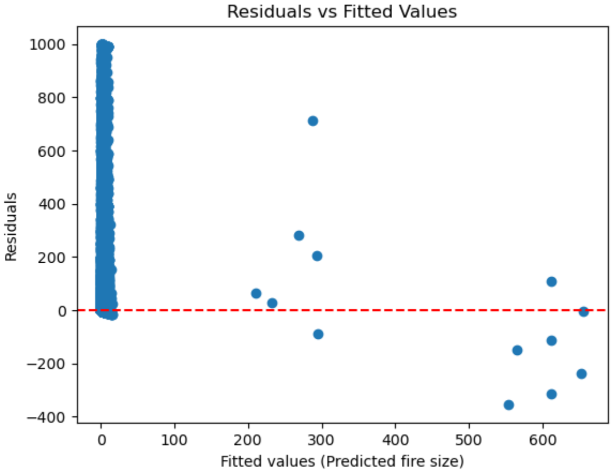

Figure 5.3: GLM with Log Transformation Residual Plot

It is noticeable that despite the transformation, the GLM's assumptions were still violated, as evidenced by the Q-Q plot (Figure 5.1), the histogram of residuals (Figure 5.2), and the residuals vs. fitted values plot (Figure 5.3). The Q-Q plot shows that the residuals do not follow a normal distribution, the histogram exhibits skewness, and the residuals vs. fitted values plot reveals heteroscedasticity. These issues suggest that the GLM model may not be suitable for predicting small fire sizes, even after applying the log transformation. This could be due to the fact that GLM assumes a linear relationship between the predictors and the target variable, which may not hold true in this case. The log transformation does not address the underlying non-linearity, which could explain why it did not improve the GLM's performance.

The MLP model showed similar performance to GLM, with RMSE increasing from 52.70 to 54.38 after the log transformation. This might be due to the fact that MLPs, being neural networks, are already capable of learning complex non-linear relationships between inputs and outputs. Since the transformation only affected the target variable and not the model's architecture, the MLP model may not have benefitted from this change.

5.2 Large Fire Size Prediction

This section presents the results of predicting large fire sizes using five different machine learning models: XGBoost, CatBoost, Random Forest, Generalized Linear Model (GLM), and Multilayer Perceptron (MLP). Since large fire size data is relatively sparse compared to small fires, predicting large fire sizes presents unique challenges, including the risk of underfitting and model instability. The results presented here examine the performance of each model both before and after applying the log transformation to address the skewness and variance instability in the target variable.

5.2.1 RMSE Results for Large Fire Size Prediction

Table 5.2 summarizes the RMSE results for the five models before and after applying the log transformation:

Model	**Before Transformation (RMSE)**	**After Log Transformation (RMSE)**
XGBoost	32159.60	30373.01
CatBoost	29130.40	30523.21
Random Forest	31256.64	30362.66
GLM	29702.24	31037.39
MLP	29818.08	31957.31

Table 5.2: RMSE Results for Large Fire Size Prediction before and after Log Transformation

5.2.2 Analysis and Comparison of Results

Before applying the log transformation, all five models showed relatively high RMSE values, ranging from 29130.40 to 32159.60. These values reflect the challenge of predicting large fire sizes, especially considering the limited size of the dataset. Among these models, CatBoost achieved the lowest RMSE of 29130.40, indicating that it performed the best before the transformation.

After applying the log transformation to the target variable, the RMSE values for XGBoost and

Random Forest dropped significantly, with XGBoost achieving the lowest RMSE of 30362.66. However, this RMSE value is still larger than the RMSE for CatBoost before transformation, indicating that the log transformation did not significantly improve the prediction accuracy for large fires. This suggests that while the log transformation helped stabilize variance and improve the model's performance for some models, it was not enough to overcome the challenges of predicting large fire sizes.

In the case of GLM, the assumption violations observed in the residuals persisted even after the log transformation. The Q-Q plot (Figure 5.4) shows that the residuals do not follow a normal distribution, as they deviate from the expected straight line, especially in the tails. This is further confirmed by the histogram of residuals (Figure 5.5), which shows skewness, and the residuals vs. fitted values plot (Figure 5.6), which reveals heteroscedasticity. These issues suggest that GLM might not be fully capturing the underlying patterns in the data, indicating that it may not be suitable for predicting large fire sizes in this context.

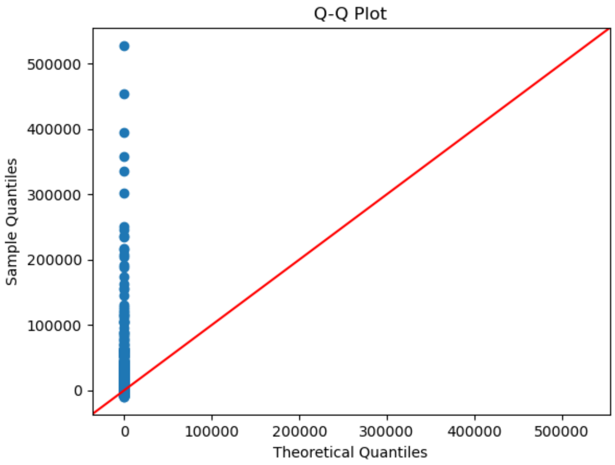

Figure 5.4: GLM with Log Transformation Q-Q Plot

5.3 Fire Size Classification Model

This section presents the results of applying classification models to predict the probability of fire sizes. The models evaluated include CatBoost, XGBoost, Random Forest, and CatBoost with

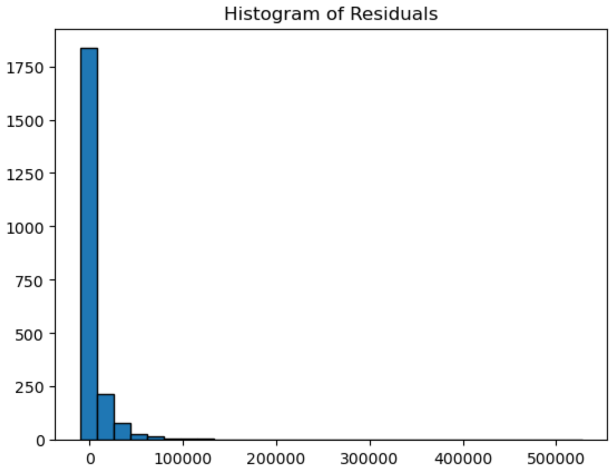

Figure 5.5: GLM with Log Transformation Residual Histogram

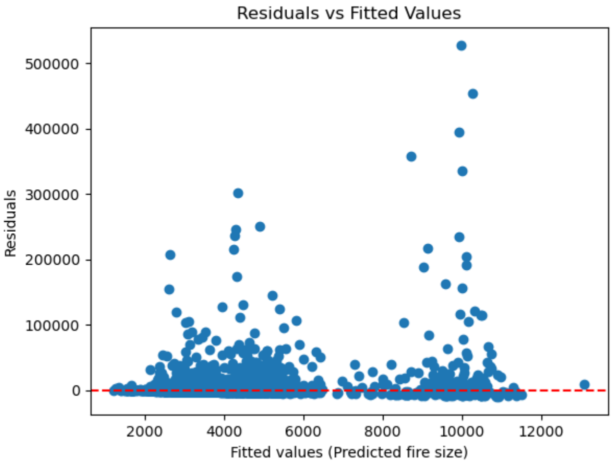

Figure 5.6: GLM with Log Transformation Residual Plot

class weights. The performance of each model is assessed using a variety of metrics, including precision, recall, F1-score, accuracy, ROC-AUC, confusion matrix, balanced accuracy, and Matthews Correlation Coefficient (MCC).

5.3.1 Classification Results for Large Fire Prediction

The classification results for the four models are summarized below.

CatBoost (without class weights)

The classification results for the CatBoost model without class weights are as follows:

Metric	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.99	1.00	1.00	373880
1	0.72	0.04	0.08	2213
Accuracy	0.99	–	–	376093
Macro avg	0.86	0.52	0.54	376093
Weighted avg	0.99	0.99	0.99	376093

Table 5.3: Classification Report for CatBoost (without class weights)

Metric	**Value**
ROC-AUC	0.92
Balanced Accuracy	0.52
Matthews Correlation Coefficient (MCC)	0.17

Table 5.4: Additional Metrics for CatBoost (without class weights)

Confusion Matrix:

$$\begin{bmatrix} 373843 & 37 \\ 2119 & 94 \end{bmatrix}$$

CatBoost (with class weights)

The classification results for the CatBoost model with class weights are as follows:

Metric	**Precision**	**Recall**	**F1-Score**	**Support**
0	1.00	0.99	0.99	373880
1	0.21	0.29	0.24	2213
Accuracy	0.99	–	–	376093
Macro avg	0.60	0.64	0.62	376093
Weighted avg	0.99	0.99	0.99	376093

Table 5.5: Classification Report for CatBoost (with class weights)

Metric	**Value**
ROC-AUC	0.93
Balanced Accuracy	0.64
Matthews Correlation Coefficient (MCC)	0.24

Table 5.6: Additional Metrics for CatBoost (with class weights)

Confusion Matrix:

$$\begin{bmatrix} 371450 & 2430 \\ 1576 & 637 \end{bmatrix}$$

XGBoost

The classification results for XGBoost are as follows:

Metric	**Precision**	**Recall**	**F1-Score**	**Support**
0	1.00	0.99	0.99	373880
1	0.19	0.29	0.23	2213
Accuracy	0.99	–	–	376093
Macro avg	0.60	0.64	0.61	376093
Weighted avg	0.99	0.99	0.99	376093

Table 5.7: Classification Report for XGBoost

Metric	**Value**
ROC-AUC	0.93
Balanced Accuracy	0.64
Matthews Correlation Coefficient (MCC)	0.23

Table 5.8: Additional Metrics for XGBoost

Confusion Matrix:

$$\begin{bmatrix} 371242 & 2638 \\ 1576 & 637 \end{bmatrix}$$

Random Forest

The classification results for Random Forest are as follows:

Metric	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.99	1.00	1.00	373880
1	0.41	0.08	0.13	2213
Accuracy	0.99	–	–	376093
Macro avg	0.70	0.54	0.56	376093
Weighted avg	0.99	0.99	0.99	376093

Table 5.9: Classification Report for Random Forest

Metric	**Value**
ROC-AUC	0.89
Balanced Accuracy	0.54
Matthews Correlation Coefficient (MCC)	0.18

Table 5.10: Additional Metrics for Random Forest

Confusion Matrix:

$$\begin{bmatrix} 373633 & 247 \\ 2043 & 170 \end{bmatrix}$$

5.3.2 ROC Curves for Large Fire Prediction

The ROC curves for each model are shown in Figure 5.7, Figure 5.8, and Figure 5.9. These curves provide insight into the power of the models in distinguishing between small and large fires. A higher AUC indicates better performance, with values approaching 1.0 indicating perfect discrimination.

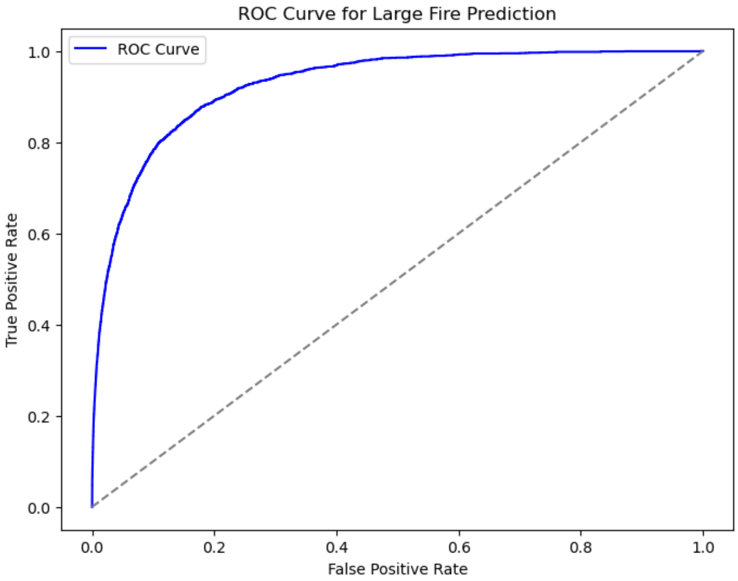

Figure 5.7: ROC Curve for CatBoost (with class weights) for Large Fire Prediction

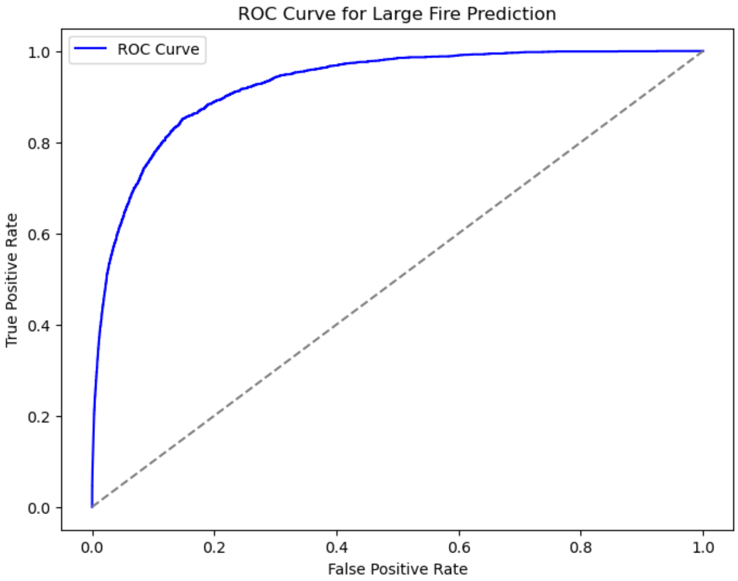

Figure 5.8: ROC Curve for XGBoost for Large Fire Prediction

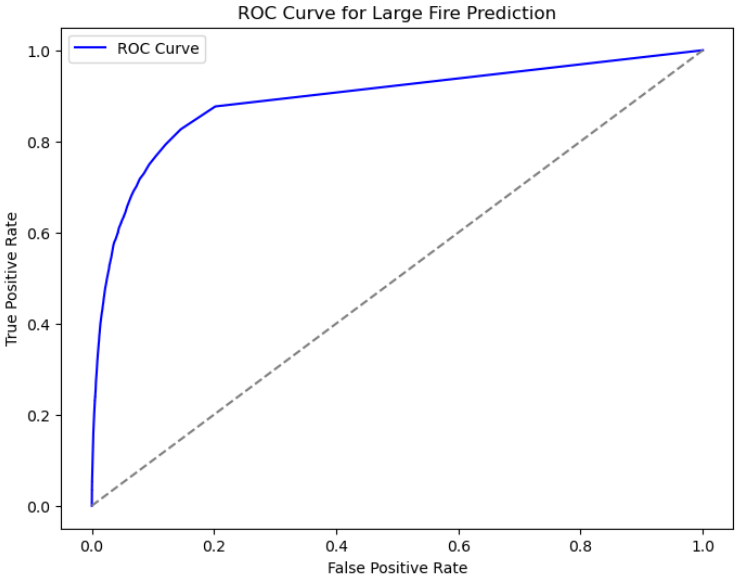

Figure 5.9: ROC Curve for Random Forest for Large Fire Prediction

5.3.3 Analysis and Comparison of Classification Models

The results demonstrate that all models achieved high accuracy, with CatBoost, XGBoost, and Random Forest performing similarly well. However, the performance metrics for each model reveal interesting insights into their strengths and weaknesses. CatBoost, even without class weights, performed well with a high ROC-AUC score of 0.92, though its recall for large fires (class 1) was relatively low at 0.04. After applying class weights, CatBoost showed improved recall for large fires (0.29) and a slight increase in balanced accuracy (0.64), which indicates a better ability to identify large fires. However, its precision for large fires remained low (0.21), reflecting a trade-off between precision and recall.

XGBoost performed similarly to CatBoost, with a ROC-AUC score of 0.93 and a balanced accuracy of 0.64. Despite achieving good overall accuracy, the recall for large fires was still low (0.29), and precision was also modest (0.19). This suggests that while XGBoost was effective at identifying small fires, it struggled to predict large fires accurately.

Random Forest, although achieving high precision and recall for small fires (class 0), struggled with large fires, with a recall of just 0.08 and a balanced accuracy of 0.54. This indicates that while

Random Forest was able to classify small fires very well, it was less effective at detecting large fires.

The Matthews Correlation Coefficient (MCC) for all models was relatively low, reflecting the imbalance in the dataset. Although CatBoost with class weights performed the best, the models still faced challenges in predicting large fires due to the class imbalance. The low recall for large fires across all models indicates that the models had difficulty identifying this minority class despite the class balancing techniques applied.

CHAPTER 6

Limitation and Future Work

This study contributes toward the identification of factors responsible for wildfire dynamics and the predictability of machine learning models in fire size forecasting. However, certain limitations about the dataset used and the models applied in this study need to be acknowledged, as these could impact the generalizability of findings and the accuracy of the predictions.

One limitation is the quality of the data itself. This dataset does not include critical features like humidity, wind speed, precipitation, topography, and land use that could potentially impact fire behavior. Furthermore, information regarding land type was not included in this data explicitly, while the land type used in this research was derived implicitly from the reporting agency. The sparsity of large fire data, especially those above 1000 acres, is still one of the biggest challenges for correct predictions.

Another limitation is that the study relies on historical data from 1992 to 2015, which may not fully capture recent trends in the behavior of wildfires influenced by climate change and evolving land management practices. This inability to reflect current fire dynamics is further limited by the lack of data after 2015. This underlines the need for updating the dataset with more recent fire incidents to better capture the ongoing shifts in the pattern of wildfires. Another limitation results from the performance of the models in predicting large fires. While the machine learning algorithms XGBoost and CatBoost performed very well in the prediction of small fires, their performance remained limited for large fires. An attempt was made at applying transformations that stabilize variance, such as the log transformation and Box-Cox transformation, but these failed to resolve the challenge at hand in accurately predicting large fires.

Moreover, the GLM model's assumptions were violated, hence leading to residual violations. For instance, GLM showed significant violations in normality and heteroscedasticity that limited its suitability for fire size prediction. These challenges illustrate the need to consider other models or hybrid approaches that can handle these challenges even more effectively, especially for imbalanced datasets with complex data structures. Limitations regarding the shortcoming of the classifier model: State-of-the-art machine learning algorithms in terms of XGBoost, CatBoost, and Random Forest were put into place within this study, in which the developed models are faced by one strong class imbalance between small and big fire classes. The algorithm detected very good performance of class in the case of the small fire while giving more inferior recall in large fire classification-the sign of minority classes, which did not develop properly in the system.

The classification of large fires could be improved by using oversampling, undersampling, or more sophisticated techniques such as anomaly detection. Besides, ensemble learning strategies could be applied to enhance the robustness of the models and, hence, their ability to predict large fires more accurately.

In the future, emphasis can be given to improving the quality of the data. More specific, continuous data for large fires can be included, such as the inclusion of other variables: real weather conditions, land management, climate change predictions. Expanding the dataset to a wider range of geography or incorporating satellite image data will most likely yield better results on wildfire dynamics and will give even better predictive models. Another direction for future work could be the improvement in the prediction of large fires by incorporating advanced machine learning techniques, such as deep learning methods or hybrid models that combine decision trees with neural networks. These methods may provide further detail in understanding fire behavior, especially those rare but catastrophic large fires. Moreover, the search for methods of causal inference may bring out the deeper relationship between environmental factors and outbreaks of fire to eventually improve management. The presence of such gaps means that, upon addressing them in the future, research will be in a position to realize more reliable and accurate wildfire predictions crucial in developing better prevention and management strategies.

CHAPTER 7

Conclusion

This study aimed to explain the dynamics in the wildfire, considering the low and high fire size with predictions of machine learning models in the determination of key variables in the occurrence of fire. Advanced models like XGBoost, CatBoost, Random Forest, and GLMs have been employed for the purposes of fire size prediction and category classification.

In all models tested, the performance on the small fire size predictions has relatively high RMSE values, and there is little significant improvement from the log transformation. Among them, XGBoost and CatBoost performed well, though XGBoost reported the lowest RMSE value both before and after the log transformation. Such consistency of this model across both transformed and non-transformed data sets underlines the robustness of this XGBoost model for predicting small fires. Therefore, it is still XGBoost that remains the most reliable for small fire size predictions without transformation since it consistently behaves across variable conditions.

This challenge was more pronounced for the large fire size prediction, since data sparsity increases with increasingly larger fire sizes. Without transformation, CatBoost worked best for large fire size predictions, but it was clear that predicting large fires remains a challenging task. While log transformation did somewhat manage to increase the performance of mainly the XGBoost and Random Forest models, it did not turn out to be sufficiently satisfactory for the models' performance to be ideal-large fire size prediction. Notwithstanding these results, certain limitations persist within the dataset itself, thereby signaling the clear need to work out further model refinements in view of higher-order exploratory techniques. Extra features can be added, and/or the use of hybrid modeling techniques that could help address these challenges in large fire predictions.

In the classification of small versus large, all models were highly sensitive in predicting small fires, although the ability to predict large ones was considerably limited by the imbalance between the classes of small and large fires.

While machine learning models demonstrated promising results in predicting small fires, significant challenges remain in predicting large fires accurately. Application of class weights increased recall in large fires, but there was more room for improvement by the model in terms of class imbalance. This may indicate that future work could apply more advanced techniques, such as oversampling, undersampling, or anomaly detection, and tune the models more carefully to take into account the imbalance between small and large fires.

Bibliography

[1] J. T. Abatzoglou and A. P. Williams. Impact of anthropogenic climate change on wildfire across western us forests. *Proceedings of the National Academy of Sciences*, 113(42):11770–11775, 2016.

[2] Giuseppe Amatulli, Fernando Peréz-Cabello, and Juan De La Riva. Mapping lightning/human-caused wildfires occurrence under ignition point location uncertainty. *Ecological Modelling*, 200(3-4):321–333, 2007.

[3] Alexandra Bjånes, Rodrigo De La Fuente, and Pablo Mena. A deep learning ensemble model for wildfire susceptibility mapping. *Ecological Informatics*, 65:101397, 2021.

[4] Kit Yan Chan, Bilal Abu-Salih, Raneem Qaddoura, Ala' M. Al-Zoubi, Vasile Palade, Duc-Son Pham, Javier Del Ser, and Khan Muhammad. Deep neural networks in the cloud: Review, applications, challenges and research directions. *Neurocomputing*, 545:126327, 2023.

[5] Esri. Wildfire size interpretation, 2016. Accessed: 2024-11-26.

[6] Adrián Jiménez-Ruano, William M. Jolly, Patrick H. Freeborn, Daniel José Vega-Nieva, Norma Angélica Monjarás-Vega, Carlos Iván Briones-Herrera, and Marcos Rodrigues. Spatial predictions of human and natural-caused wildfire likelihood across montana (usa). *Forests*, 13(8):1200, 2022.

[7] Dmitri A. Kalashnikov, John T. Abatzoglou, Paul C. Loikith, Nicholas J. Nauslar, Yianna Bekris, and Deepti Singh. Lightning-ignited wildfires in the western united states: Igni-

tion precipitation and associated environmental conditions. *Geophysical Research Letters*, 50(16):e2023GL103785, 2023.

[8] Alex W. Marden, Thoralf Meyer, and Kelley A. Crews Meyer. Regional fire occurrence in southern africa using bfast iterative break detection in seasonal and trend components of a modis time series. *South African Geographical Journal*, 105(2):200–221, 2023.

[9] National Oceanic and Atmospheric Administration. Climate data online (cdo) datasets, 2024. Accessed: 2024-11-26.

[10] Binh Thai Pham, Abolfazl Jaafari, Mohammadtaghi Avand, Nadhir Al-Ansari, Tran Dinh Du, Hoang Phan Hai Yen, Tran Van Phong, et al. Performance evaluation of machine learning methods for forest fire modeling and prediction. *Symmetry*, 12(6):1022, 2020.

[11] Volker C. Radeloff, David P. Helmers, H. Anu Kramer, Miranda H. Mockrin, Patricia M. Alexander, Avi Bar-Massada, Van Butsic, Todd J. Hawbaker, Sebastián Martinuzzi, Alexandra D. Syphard, and Susan I. Stewart. Rapid growth of the us wildland-urban interface raises wildfire risk. *Proceedings of the National Academy of Sciences*, 115(13):3314–3319, 2018.

[12] Krishna Rao, A. Park Williams, Noah S. Diffenbaugh, Marta Yebra, Colleen Bryant, and Alexandra G. Konings. Dry live fuels increase the likelihood of lightning-caused fires. *Geophysical Research Letters*, 50(15):e2022GL100975, 2023.

[13] Marcos Rodrigues and Juan De La Riva. An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software*, 57:192–201, 2014.

[14] Ignacio San-Miguel, Nicholas C. Coops, Raphaël D. Chavardès, David W. Andison, and Paul D. Pickell. What controls fire spatial patterns? predictability of fire characteristics in the canadian boreal plains ecozone. *Ecosphere*, 11(1):e02985, 2020.

[15] Rachael Tatman. 1.88 million us wildfires, 2019. Updated: 5 years ago, Accessed: 2024-11-26.

[16] Daniel Vecín-Arias, Fernando Castedo-Dorado, Celestino Ordóñez, and José Ramón Rodríguez-Pérez. Biophysical and lightning characteristics drive lightning-induced fire occurrence in the central plateau of the iberian peninsula. *Agricultural and Forest Meteorology*, 225:36–47, 2016.